DOCUMENT RESUME

ED 364 232 IR 054 720

AUTHOR	Eaton, Nancy L.; Andre, Pamela Q. J.
TITLE	The National Agricultural Text Digitizing Project: Toward the Electronic Library. Report of the Pilot Project, Phases 1-2, 1986-1992.
INSTITUTION	Iowa State Univ. of Science and Technology, Ames.; National Agricultural Library (DOA), Washington, D.C.
SPONS AGENCY	Department of Education, Washington, DC.
PUB DATE	15 Nov 92
NOTE	106p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Academic Libraries; Access to Information; Acid Rain; Administrators; Cooperative Programs; Financial Support; Higher Education; *Information Dissemination; Information Networks; Land Grant Universities; *Library Services; Library Surveys; National Libraries; *Optical Data Disks; *Optical Scanners; Research Projects; Shared Library Resources; Special Libraries; Technological Advancement
IDENTIFIERS	Aquaculture; *Digitizing; *National Agricultural Library MD; Text Handling

ABSTRACT

The National Agricultural Text Digitizing Project (NATDP) began in 1986 with cooperation between the National Agricultural Library and the University of Vermont, and then expanded to include 45 land-grant university libraries and 1 special library. The first activity was to evaluate the new technology of optical scanning. The project was designed to test the feasibility, cost, and effectiveness of newly emerging technologies for capturing page images, providing access to their content, and disseminating them for use in agricultural communities. Four CD-ROM sets, covering aquaculture, international agricultural research, Agent Orange, and acid rain, were produced by scanning printed materials and were distributed to the universities for use and evaluation. Overall, managers concluded that scanning and text digitizing are effective technologies for disseminating information. A few cautions were noted. A great deal of time was required to participate, and participants needed to be able to provide the technical and professional support required. Managers recommended continuation of the project, with specific improvements, and suggestions for particular products. The cooperation of the various institutions and their funding support were essential in program success. Nine tables and three diagrams illustrate the discussion. Nine appendixes provide supplemental information, including additional reports about program operation. A nine-item bibliography of articles about the NATDP is included. (SLD)

ED 364 232

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

National Agricultural Text Digitizing Project:

Toward the Electronic Library

1992

BEST COPY AVAILABLE

IR054720

The National Agricultural Text Digitizing Project: Toward the Electronic Library

Report of the Pilot Project,
Phases 1-2
1986 - 1992

Nancy L. Eaton
Dean of Library Services
Iowa State University
of Science and Technology

and

Pamela Q.J. André
Associate Director
Information Systems Division
National Agricultural Library

November 15, 1992

Executive Summary

The National Agricultural Text Digitizing Project (NATDP) began in 1986 with cooperation between the National Agricultural Library (NAL) and the University of Vermont, but grew in the next four years to include forty-five land-grant university libraries and one special library. The first activity was to evaluate the new technology of optical scanning, a technology that allows printed text and images to be captured in digital form for publication in electronic form.

Project participants were anxious to determine optical scanning's application for improving worldwide access to agricultural literature. Specifically, the purpose of the project was to test the "feasibility, cost and effectiveness of newly emerging technologies for capturing page images, providing access to their content and disseminating them for use in the agricultural community." During the years of the project, four CD-ROM sets were produced by scanning printed materials and provided to land-grant university libraries for use and evaluation. The CD-ROMs covered aquaculture, international agricultural research, Agent Orange, and acid rain.

The cooperative nature of the NATDP, including the varied funding support received, was a key reason for its success. The project was funded through NAL's small research and development budget and by contributions from land-grant university libraries. Funds were also provided by the Science and Education Office of the U.S. Department of Agriculture and through a grant to the University of Vermont from the U.S. Department of Education.

Overall, NATDP managers concluded that scanning and text digitizing are effective technologies that work in providing agricultural information more readily to those throughout the nation who need it. This conclusion was tempered, however, with several caveats. First, a great deal of time is required to participate in such experiments. It is critical that participants be willing to spend the time required. Also, participants must be able to provide the technical and professional support such projects require. Lack of technical expertise prevented several test sites from finishing the installation of equipment and software. Next, users of the technology should be aware that upgrades most likely will be needed. Hardware and software evolved rapidly during the period of the research and demonstration project. Finally, parts of this technology are complex. The full text formatting and searching phase of the project illustrates this.

Based on these conclusions, the project managers recommended that NAL should continue a National Agricultural Text Digitizing Program. More specific recommendations were also made, including:

- Make page images available for display on any monitor
- Upgrade the text recognition software
- Hire permanent staff
- Develop a process for selecting materials for scanning
- Determine whether ASCII-only, ASCII plus bit-mapped images, or bit-mapped images with fixed field retrieval elements will be used
- Determine specifications for archiving data so that content is independent of the workstation platform and the specific distribution medium

- Monitor NAL's in-house scanning capabilities to ensure that they keep up with those being developed in the commercial sector
- Develop a system to allow users to gain access to materials from personal computers through networks
- Pursue licensing solutions that allow networked access to CD-ROM collections
- Ensure the integrity of data and content regardless of hardware and software changes
- Protect and refresh master disks or source databases as the electronic storage medium ages
- Ensure that NAL uses software that supports multiple platforms (DOS, OS/2, Macintosh, UNIX operating systems) since the user base of hardware must be used for distribution

In 1992, NATDP became a fully operating "program" through which NAL began routinely to produce CD-ROMs containing a variety of agricultural information and making these available to land-grant university libraries. The disks contain full text and images which are accessed through the retrieval package Windows Personal Librarian, developed by Personal Library Software, Inc. of Rockville, Maryland. As funds become available, NAL is moving ahead in producing other CD-ROMs containing portions of its collections.

With assistance from the American Society of Agronomy, NAL has produced a CD-ROM containing sixteen volumes of that society's journal. Hoping to build on this cooperation with a professional scientific society, NAL has sponsored a workshop for the editors and publishers of other major agricultural professional societies to explain the value of the text digitizing program in increasing the accessibility and preservation of the societies' journals. NAL also has worked with Tuskegee Institute to produce a disk containing selected materials from the collection of works by famed scientist George Washington Carver. This activity illustrated how microfilm can be converted to electronic page images that can be accessed electronically.

As the nation's chief agricultural information resource, NAL is continually working to expand its user community and its information services to that community. With an ongoing program to develop electronic agricultural information products, NAL is clearly moving into the electronic future and will continue to provide leadership to the national agricultural information community.

Acknowledgements

This project could not have been done without the tremendous efforts of staffs and local participants at the forty-six participating libraries. Particular thanks go to the staffs of the three lead institutions: The University of Vermont, Iowa State University, and the National Agricultural Library. The principal investigator (Pamela Q. J. André) and the National Project Manager (Nancy L. Eaton) could not have succeeded in carrying out this massive project without the support and dedication of these individuals.

At the National Agricultural Library, Judith Zidar served as Project Coordinator for the Central Scanning Facility, with technical support provided by Gary McCone, John Stetka, and Dan Starr. Carol J. Shore provided support with the statistical analysis for the study of scanning error rates. Staff of the Aquaculture Information Center provided support for the development of the CD-ROM on aquaculture. The Educational Programs Unit assisted with training and demonstrations for NATDP products.

The University of Vermont (UVM) served as the initial cooperating institution with NAL. Jerry V. Caswell and Albert Joy served as local project coordinators, with support from staff of the Reference Department who helped draft and pre-test field survey instruments.

Iowa State University assumed the role of cooperating institution when Nancy Eaton accepted a new position at that institution in 1989. Responsibility for the field testing was then assumed by Diana Shonrock and Gerard McKiernan; once again the reference department aided in pre-testing.

The Executive Secretariat of the Consultative Group for International Agricultural Research (CGIAR) cooperated with the NATDP in the development and distribution of disk number two, *Food, Agriculture and Science*. Eleanor Frierson served as coordinator and liaison for this aspect of the project.

Clifford A. Lynch, Director, Division of Library Automation, University of California System, served as technical advisor throughout the six years of this project. He was particularly helpful with advice on technical platforms and technical standards.

We are also appreciative of the strong support from the directors of the land-grant libraries who provided matching funds and advice throughout the project. Special thanks go to those who served on the technical advisory panel during the six years of the project: John Beecher, H. Joanne Harrar, Noreen Alldredge, Paul Gherman, Tom Shaughnessy, and Charlene Mason.

Foreword

A Technology Perspective on the National Agriculture Library's Text Digitizing Project

by

Clifford A. Lynch
Director, Division of Library Automation
University of California System

Introduction

The NATDP presented several unusual technical challenges that may not be fully evident to many readers of this report. I believe that it illustrates some problems that are becoming increasingly common as we try to manage technology-intensive projects. This foreword outlines briefly some of the issues, how NATDP addresses them, and the lessons learned. And there were real lessons learned in at least three technology-related areas: the difficulties of managing a leading-edge technology intensive project in an environment where the product life cycle is measured in periods of two years or less; the complexities of realizing an image technology application in actual working hardware and software, including a base of workstations at user sites over which the project had relatively little direct control; and the problems of coping with an environment where relevant standards were at best immature and those that formally existed were not implemented in marketplace products.

Managing Short-Life Cycle Technology

Recall first that the two phases of the NATDP covered in this report span 1986 to 1992—six years. This is easily three generations of workstation hardware and two generations of software. Further, a number of hardware components such as memory and high-resolution bit-mapped displays, which are critical for imaging applications, have become much less expensive over these six years.

The primary objective of the NATDP was not to create new technology, but to employ a developing set of technologies early in their commercial life cycle to understand how effectively they worked for the NATDP applications. Experience gained from the NATDP would help plan future operational production activities. Thus, NATDP invested early in scanning technology and display technology using Science Applications International Corporation (SAIC) as a system integrator. This was an expensive route (indeed, absurdly so in terms of today's prices for off-the-shelf hardware and software technology), but it allowed the project to gain experience prior to the general availability of these technologies.

One substantial problem was the change in technology even during the life of the project. A key goal of the project was to deliver to participating libraries a series of CD-ROMs for evaluation of their software and content. At the start of the project, a basic platform was defined and libraries installed it. By the end of the test period, this basic hardware configuration was clearly obsolete, and we faced the choice requiring a more elaborate platform as a minimal configuration for the final CD-ROMs, such as a 386 machine with Windows, or a slower, smaller DOS machine. I think we learned that, in developing experimental systems that push the technology base, one should target the most expensive hardware and software platform within reach, recognizing that when one is ready to go into large-scale deployment, such a configuration will be commonplace.

Another difficult issue in managing the technology base was accommodating the diversity of platforms. Ideally, we should have developed a set of experimental CD-ROMs targeted for multiple platforms, since realistically the production environment contains a considerable diversity of platforms—not only various types of IBM PCs, but Apple Macintoshes and UNIX workstations. This would not only have introduced enormous and possibly insurmountable complexity into the evaluation process (since the characteristics of the underlying hardware and operating system would likely color patrons' views of the overall use ability of the system), but also would have added an additional layer of technical complexity simply in getting anything to work. In 1988, multi-platform applications on CD-ROM were much more theory than practice, particularly if they involved imaging. Practically speaking, they are only becoming feasible today with the widespread adoption of the ISO 9660 CD-ROM directory format, the availability of multi-platform CD-ROM retrieval software, and the widespread implementation of at least some de facto standards for image format and compression.

There were some interesting, rather unforeseen implications to our decision to standardize the hardware and operating systems platforms for the duration of the experiments. One was that some sites with little experience with the IBM PC and Windows had difficulty installing the hardware and software and getting it running. Other sites with a lot of expertise in this configuration had an easier time. The importance of local site expertise in making a project like this work cannot be underestimated, and, in some cases, standardizing the platform prevented us from fully exploiting this local expertise. As the production system moves to a multi-platform distribution approach, this will be less and less of a problem, and, indeed, this is one of the key justifications for undertaking a multi-platform direction in the production system.

There are also certain aspects of the underlying technology used in the project that still surprise me with their deficiencies—such as the available optical character recognition (OCR) technology. It would seem that one could do much better by employing dictionaries, databases of typical errors to help in automated correction, and perhaps multiple OCR algorithms. But such technologies were not really on the marketplace when the project procured its OCR technology, and the purpose of the project was to test available technology, not to develop new technology. Thus, such opportunities were not explored within the scope of the project, no matter how promising they might have seemed.

It is also important to remember that software technology evolved a great deal during the six years of the project. In 1987, the choices for image-capable CD-ROM retrieval software were limited, and in many ways all choices were unsatisfactory. As the project report indicates, by 1990 there were many more options. But the project called for early initial testing; we did not have the option of waiting for the ideal CD-ROM retrieval package. The project had to balance constantly

8

the need to deploy test systems from which to learn with the temptation to wait another year for improved technology and marketplace offerings. This was a central dilemma: Deploy too soon, and the flaws of immature software and hardware would render the findings suspect; wait too long and lose the opportunity to make decisions about the production system based on actual data gathered from field tests of prototypes. In my opinion, the project successfully balanced these two imperatives, though it may sometimes be hard to realize why certain choices were made unless one remembers the very limited alternatives available then.

Imaging Systems

It has been clear since the early 1980s that electronic imaging technologies could have an enormous effect on library automation and information access. A number of authors (including myself) have written rather glibly about these potentials. In fact, if we learned anything from this project, it is that integrating a robust, working production system that captures, manages, and publishes images is hard!

I believe that the report accurately conveys the difficulties and frustrations of getting a reliable image capture and management system into production operation at NAL. Again, this has become much easier in the last couple of years with the introduction of a wide range of more user-friendly commercial products. But my experience is that it is still more difficult to develop image applications than many realize. Also, image applications tax hardware and software systems to the limit. Many people are surprised that image capture, storage, retrieval, and display are far from instantaneous, even on high-end platforms. This is an important lesson to remember as the third phase of the NATDP explores networked image delivery, where presentation delays are greater and more variable. It will be some time before information technology can make image retrieval and display as quick as we would like it to be.

The decision to deploy the "high-end" SAIC image display workstation was particularly troublesome. We did not like this system on several counts: It was vastly expensive; it was unique; and it did not employ any of the relevant emerging standards with which we would have liked to gain experience. But the decision to deploy a limited number of these systems was ultimately pragmatic: It worked; it was available; it integrated with the image capture system; and it was reasonably fast. It is only now that standard, off-the-shelf systems offer even minimally satisfactory performance in imaging applications. Arguably, from a hardware point of view, we might have been better served by selecting some type of reasonably expensive UNIX workstation, as these offered bit-mapped display support of reasonable quality earlier than the PC world. There was essentially no CD-ROM retrieval software available for this platform, and little support expertise at the participant sites. Thus, at best, we would have traded one set of problems for another.

Another important, and frequently underestimated issue with imaging applications is the need for constant quality control. A scanning system will not automatically produce a continuous stream of perfect-quality images. A great deal of management attention needs to be focused on ensuring that the images are clear, contrast is correct, and imaging artifacts are kept to a minimum. This was another difficult lesson. The project also investigated and ultimately debunked a number of widely held assumptions about the ease with which OCR technology could be used to produce good quality ASCII versions of printed text that had been imaged. Here again, a substantial degree of caution about quality control issues proved to be a wise approach.

Standards Issues

The issue of standards pervades the entire project. At the end of 1989, I prepared a document on future platform considerations for the NATDP (see Appendix 4), and it is interesting to see the areas in which this missed the mark. While I believe that the basic thrust of the document is still correct, it did not foresee the widespread acceptance of Microsoft Windows in the PC marketplace and the rather slow penetration of IBM's OS/2 system. It did not even consider the new systems, such as Microsoft's NT, which are now serious concerns in planning for systems that will run on a future installed base of workstations. The basic point made in this document—that the X Window system remains the only real platform-independent means of handling bit-mapped displays—is still true. But for real imaging applications, X remains problematic and there is active work underway to develop a set of functions called the X Imaging Extensions which will be essential to achieving high-performance, high-quality, platform-independent imaging applications in a networked environment.

ISO 9660 for CD-ROMs has proven to be a viable, widely accepted standard. But it has become clear that it has some problems in multi-platform applications, particularly if UNIX as well as the PC and the Macintosh are to be supported, and work is underway to make the necessary extensions to the standard.

Perhaps the two most vexing issues, and those on which progress has been slowest, have been in the area of image formats and network-wide digital object IDs. Only this year has the Internet Engineering Task Force (IETF) published RFC 1314, which describes a format for the network transmission of monochrome bit-mapped images based on a specialization of the Tagged Image File Format (TIFF). Hopefully, this will become a de facto standard and will see wide adoption in the next year or so. Realistically, this standard is probably our current best hope for an archival format: It is both well-defined and simple, and it seems likely that images stored in this format could be algorithmically converted to any future image transmission format. But RFC 1314 does not really address the larger structure of an imaged article as a collection of images that can be individually retrieved. (Though it does support multi-page images, this support is really not well-adapted to random access to pages within the article.) Considerable research and experimentation will provide the necessary basis to standardize such a higher-level format. Work underway at institutions including Carnegie-Mellon (Project Mercury), Cornell (CORE and various digital preservation projects), and Yale (again, digital preservation) should provide valuable insight for such development. The more complex issues of color images and compound documents are not addressed by any standards that have been well-tested in implementation and have seen wide adoption in generally available products. Monochrome bit-mapped images remain very much state of the art.

Finally, digital object IDs will be needed to establish the link between abstracting and indexing (A&I) records such as those generated at NAL and the images produced by projects like NATDP. Currently, this is being handled by an ad hoc method in the NATDP project since there is no standard. But, as these applications move into a network environment and have to survive across multiple generations of technology (consider, for example, regenerating the image databases onto the successor technology to CD-ROM while still retaining the A&I records from AGRICOLA as an access apparatus), standards will be needed. Currently, an IETF working group is developing proposals in this area. They are badly needed and should, in the long term, meet the needs of the production program in text digitizing.

Conclusions

It has been a privilege and an education to serve as Technical Consultant to the NATDP for the past six years. While others have been speculating about potential applications of imaging technology for preservation and information distribution, the NATDP has actually been accumulating an enormously valuable base of actual experience and data about the practical issues of using these technologies. It is certainly true that today we could do much of what the NATDP did in 1986 for less money. But pioneering is never without cost, and I believe that the broader community will benefit greatly by the hard-won knowledge gained by the NATDP. It is already clear that the project has been successful in several dimensions. It has generated continued production activity at NAL in distributing agricultural information on CD-ROM, which has benefitted greatly from the NATDP prototypes. Perhaps more importantly for the long term, NAL has already accumulated a significant database of images of agricultural literature which can be reused in a wide range of applications, the first of which is the networked delivery prototype being explored in Phase 3 of the project. And, I believe, the project has generated a valuable understanding of issues involved in technology management and the need for standards that will be applicable to a wide range of future projects that explore the use of information technology in library automation and information distribution.

September 21, 1992

Introduction

In 1986 the National Agricultural Library (NAL) began a cooperative project to evaluate optical scanning technology as a method of capturing text and images in digital form for publication on CD-ROM. Over the next four years participation in the project grew to include forty-five land-grant universities and one special library. (See Appendix 1 for participants.)

The purpose of the project was to test the feasibility, cost, and effectiveness of newly emerging technologies for capturing page images, providing access to their content, and disseminating them for use in the agricultural community. This was to be done through the use of a turnkey scanning system which would help determine whether it was possible to provide in-depth access to agricultural literature while preserving it from deterioration. A panel of land-grant library directors was brought to NAL for two days in September 1986 to see demonstrations of prototype equipment being tested at the Smithsonian Institution. The panel advised NAL's director, Joseph Howard, that the technology was of sufficient interest that it warranted further investigation. Thus, NAL mounted a pilot project to test the applicability of this technology for storage, dissemination, and preservation of agricultural information.

From 1986 through August 1989, the pilot project was managed jointly by NAL and the University of Vermont. From August 1989 through June 1992, the completion of Phase 2, the project was managed jointly by NAL and Iowa State University. Pamela Q.J. André, Associate Director for Automation at NAL, served as principal investigator and Nancy L. Eaton, Director of Libraries at the University of Vermont (and subsequently Dean of Library Services, Iowa State University), served as project director.

A technical advisory panel was established to guide the effort. Panel members included John Beecher, North Dakota State University; H. JoAnne Harrar, University of Maryland; Noreen Alldredge, Montana State University; Paul Gherman, Virginia Polytechnic and State University; and Tom Shaughnessy, University of Missouri, who was succeeded in 1990 by Charlene Mason, Director of Library Systems, University of Minnesota. Dr. Clifford Lynch, Director of Library Automation for the University of California Office of the President, served as technical consultant. The panel met twice a year at the midwinter and annual meetings of the American Library Association to review progress and advise on the project.

Project management was vested in the University of Vermont (and later Iowa State University) through cooperative agreements between those universities and NAL based on the mutual interest of the institutions in exploring new technologies for sharing and accessing library materials. Through these agreements, the University of Vermont and Iowa State University were able to coordinate the planning and operation of the pilot project, coordinate the participating land-grant libraries, coordinate and manage the diverse project funding and vendor contracts, and oversee the field tests and evaluations. NAL housed and staffed the central scanning and production of disks and provided technical support to the participating land-grant libraries.

Project funding came from a number of sources. A small amount was made available from the research and development funds of NAL as seed money. Land-grant libraries were invited to contribute $3,000 a year for two years to participate in the project and to receive the CD-ROM disks produced by the project; this was extremely important when additional funds were sought from other sources, as it demonstrated the land-grant libraries' commitment to the project. It is the cooperative nature of the project, including the funding, which has been the basis for its success. Additional funding was provided from the United States Department of Agriculture Assistant Secretary for Science and Education through the Evaluation Studies Program. Phase 2 was conducted by the University of Vermont as a "project within a project" under a U.S. Department of Education Higher Education Act Title II-C grant. Table 1 summarizes funding sources and amounts. Table 2 summarizes project expenditures.

Table 1

Sources of Funds

University of Georgia Indexing Funds (transfer)	$ 50,000
NAL/USDA	620,000
Land-grant libraries	230,000
HEA Title II-C Grant to University of Vermont	200,000
Total Funding, Phases 1-2	$1,100,000

Table 2

Summary of Expenditures

Phase 1	
Systems integration (SAIC)	$ 492,174
Equipment (excluding SAIC equipment)	82,699
Editing of scanned material	69,128
Retrieval software licenses	46,483
Travel and planning meetings	15,455
Technical consulting	14,328
Field evaluation	18,767
Software evaluation	17,706
Miscellaneous	17,410
Phase 2	
Acid Rain collection	230,480
Technical consulting	1,000
Phase 3	
North Carolina State University subcontract for software development	10,000
NAL Production System	
Kansas State University subcontract	81,000
Total Expenditures	$1,100,000

Objectives of the Pilot Project

Vendors of digitizing technology were marketing digitizing equipment as off-the-shelf hardware and software that could be used with little local technical support or user experience. Such systems typically were used in business or government environments where material to be scanned was fairly uniform and predictable. However, the library environment is a particularly difficult one in that library materials vary widely in age, quality of print and design, and variety of type fonts. NAL staff and the land-grant library directors wanted a research and demonstration project to test out the viability of the technology with difficult-to-scan library materials and with library patrons as users of the retrieval software. Some of the specific questions for which evidence and experience was sought included:

1. Storage capacities of a CD-ROM disk which held software, bit-mapped images, and ASCII text.

2. Acceptance by users of bit-mapped images with fixed field subject retrieval versus key word and Boolean retrieval on full text but without access to the bit-mapped graphics, or a combination of bit-mapped images for page replication and ASCII for key word and Boolean retrieval.

3. Effort and staffing required to set up and operate a centralized scanning facility.

4. Need for standards for bit-mapped image files for storage and preservation of the scanned data.

5. Comparison of software retrieval packages available for manipulation of full text with associated images.

6. Error rates and editing needed to produce acceptable data when scanning the variety of material in a library.

7. Workstation platforms needed for effective use of electronic publications. Put another way, what were the platform assumptions and software requirements to produce an acceptable product, and did software exist that met these requirements?

Project Description

The purpose of the project was to test the feasibility, costs, and effectiveness of newly emerging technologies for capturing page images, providing access to their content, and disseminating them for evaluation to the agricultural community. The system was to use an optical scanning methodology which would convert full text, including graphics and illustrations, to a digitized page image and then process the image to convert the text to digitized ASCII form. This digitized material could then be stored on a variety of electronic devices for dissemination. For this project, CD-ROM was the distribution medium.

The project was organized into three phases. Phase I, the pilot project, tested an optical scanning system and a number of indexing/search software systems with a variety of high-visibility agricultural materials. The intention was to use vendor-supplied, off-the-shelf scanning hardware and software; and a contract was signed with Science Applications International Corporation (SAIC) as system integrator for the Central Scanning Facility's equipment and software.

The first collection to be scanned was on aquaculture. Four thousand pages of the most important, non-copyrighted aquaculture material were scanned and digitized. Both the page images and the ASCII text were mastered onto a CD-ROM using the TextWare software by UNIBASE.

A second CD-ROM was created in conjunction with the Consultative Group on International Agricultural Research (CGIAR). The material for this disk consisted of the most important papers on international agricultural research as determined by the CGIAR centers. CGIAR is an association dedicated to supporting a system of agricultural research centers around the world. It is supported by the World Bank and the United Nations. This disk uses the KAware2 retrieval software by Knowledge Access International, Inc.

A third CD-ROM disk was developed on the topic of Agent Orange. This was a very important historical collection on the creation and use of this chemical defoliant. The collection currently resides in boxes at NAL due to lack of funds for processing. This project provided an opportunity to make some part of the collection available. The software used for this disk was Personal Librarian by Personal Library Software, Inc. (PLS).

The pilot project was structured with two levels of workstations, with the high end workstations recognizing the proprietary nature of high-resolution image handling. Five test sites installed high-resolution workstations with the capability to display and print page images which were as readable as the originals. The remaining sites installed standard microcomputer workstations which could view text only as ASCII text files. The reason for this difference was the high cost of the high-resolution workstations. Project leaders wanted to test whether users would accept the lower quality displays, since the cost of the workstations was significantly lower than the full-image stations and since there was such a large base of the low-end microcomputers already in the field.

In Phase II of the project, the University of Vermont completed an in-depth project on acid rain materials. Utilizing funding from a U.S. Department of Education HEA Title II-C grant, this aspect of the project used a vendor to capture text and image data from a major collection of Canadian research materials on acid rain. Saztec was selected to convert the text to machine-readable format because the Central Scanning Facility at NAL was behind schedule due to early problems with installing the original scanning workstations and could not meet the project deadlines of the grant. It was, however, also a way to compare keying of data to the scanning experience. This two-disk set was accessed using the retrieval package KAware2 by Knowledge Access International, Inc., which included vendor enhancements to the software based on field test experience with the CGIAR disk using an earlier version of the software.

An evaluation of field experience with each CD-ROM distributed as part of the pilot project has been completed by Iowa State University and is described later in this report.

Phase III of the project focused on alternative delivery mechanisms using the developing national network infrastructure. Dr. Clifford Lynch, Director of Library Automation for the University of California System, completed a state-of-the-art survey of telecommunications options for transmission of digital data. Based upon the results of this report, NAL and North Carolina State University have begun a project to test the feasibility of transmitting document images over the Internet. This report summarizes the results of Phases 1-2. Phase 3, which has been funded by two HEA Title II-D grants awarded to North Carolina State University, will be reported on separately at a later date. Appendix 8 includes a brief description of Phase III.

Central Scanning Facility

Scanning Equipment and Workflow

The Central Scanning Facility was established at NAL in January 1988, with the installation of the scanning workstation by SAIC. Shortly thereafter, a second workstation was installed to be used for editing text and testing database retrieval. Diagram 1 summarizes the logical steps in the scanning process. Diagram 2 illustrates the purpose and sequence of equipment used. Diagram 3 summarizes the final workflow utilized by the NAL Central Scanning Facility staff.

The initial configuration of the scanning system, which closely mimicked the system designed by the Smithsonian Institution, was as follows:

Hardware

- NEC PowerMate II 286 Microcomputer, with 640K RAM, 230Mb hard disk, 1.2Mb floppy drive
- Ricoh High Speed Optical Scanner
- Calera RS 3000 Recognition Server
- LaserView 17" High-Resolution Portrait Monitor
- LaserData Compression/Decompression Processor
- Ricoh 4081 Laser Printer
- Maxtor 5-1/4" WORM Drive with caching disk
- Alloy PC 9-Track Tape Drive

Software

- MS-DOS 3.1 Operating System
- Calera Scanning and Text Recognition Software
- LaserView Image Handling Software from LaserData
- LaserBank WORM Management Software from Atlantis Research
- SAIC Management and Tracking System (customized menu-driven operations software)

The editing/retrieval workstation consisted of the following:

Hardware

- NEC PowerMate II 286 Microcomputer, with 640K RAM,

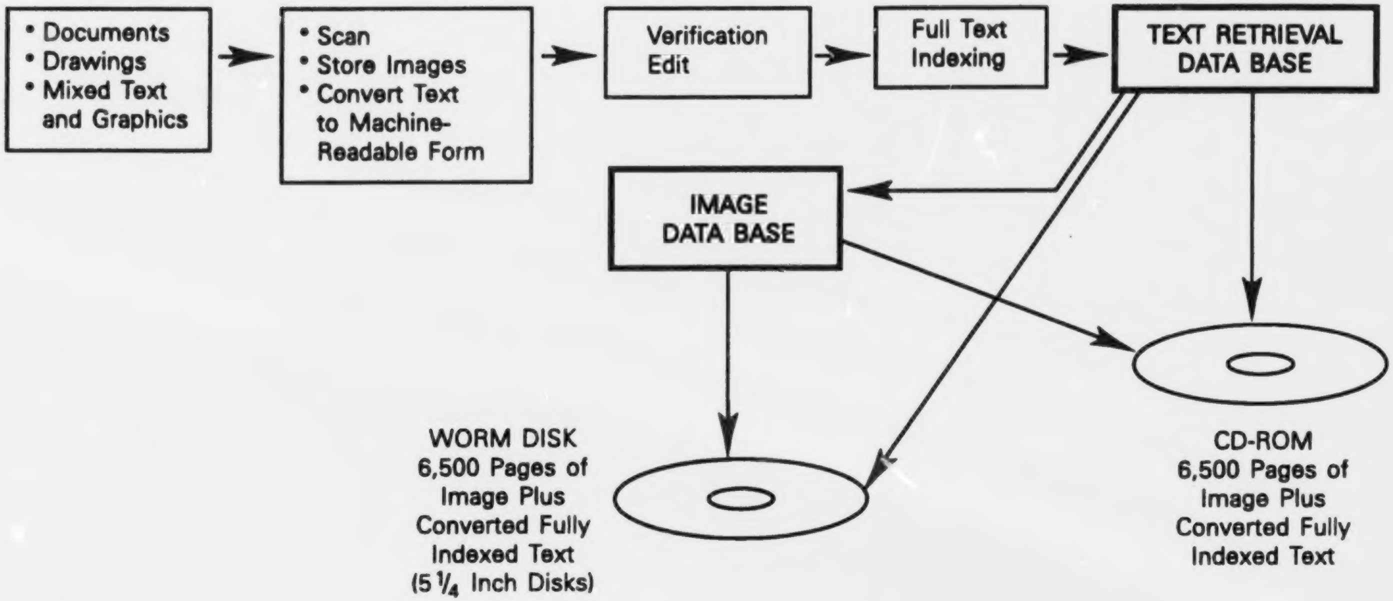
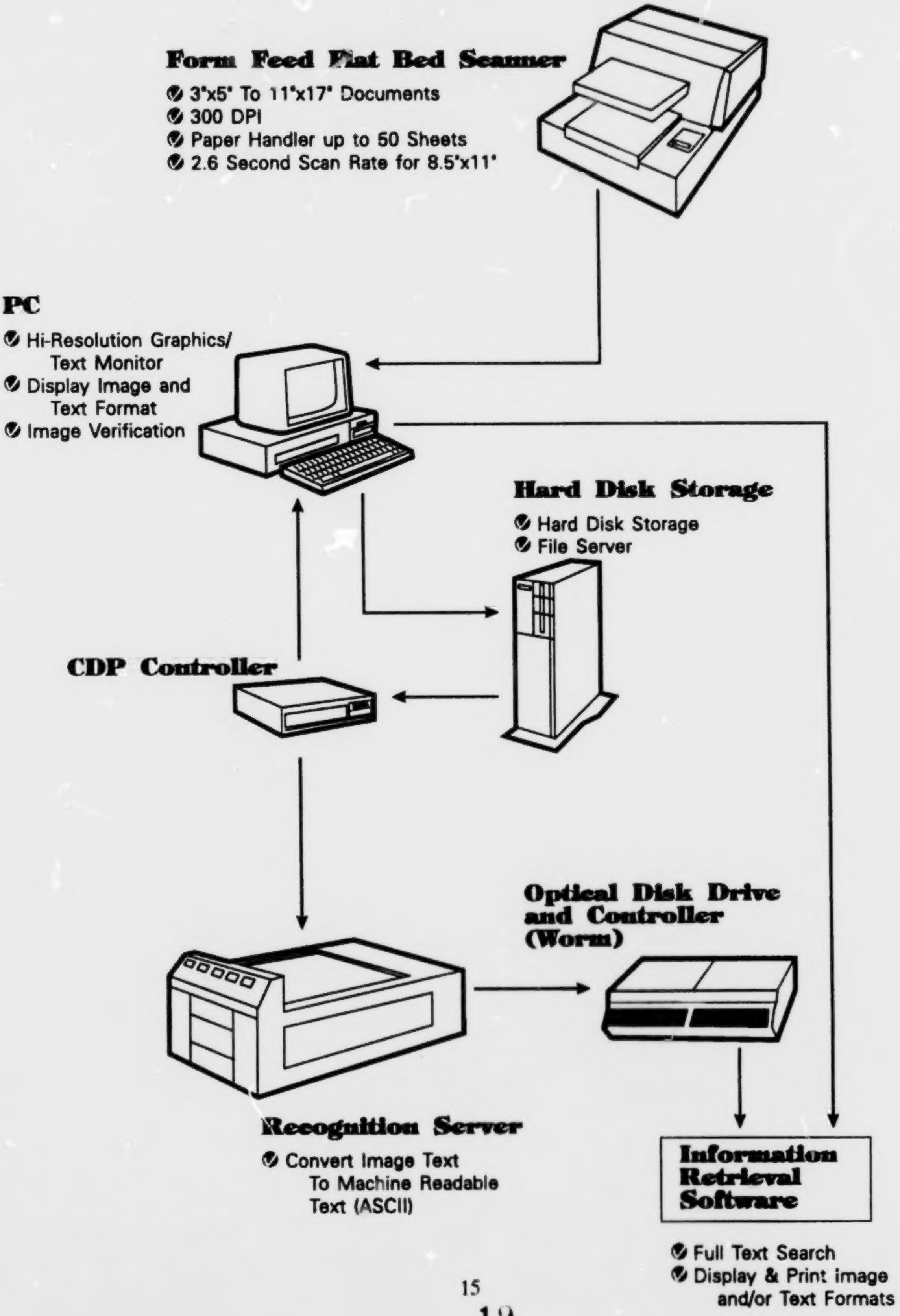
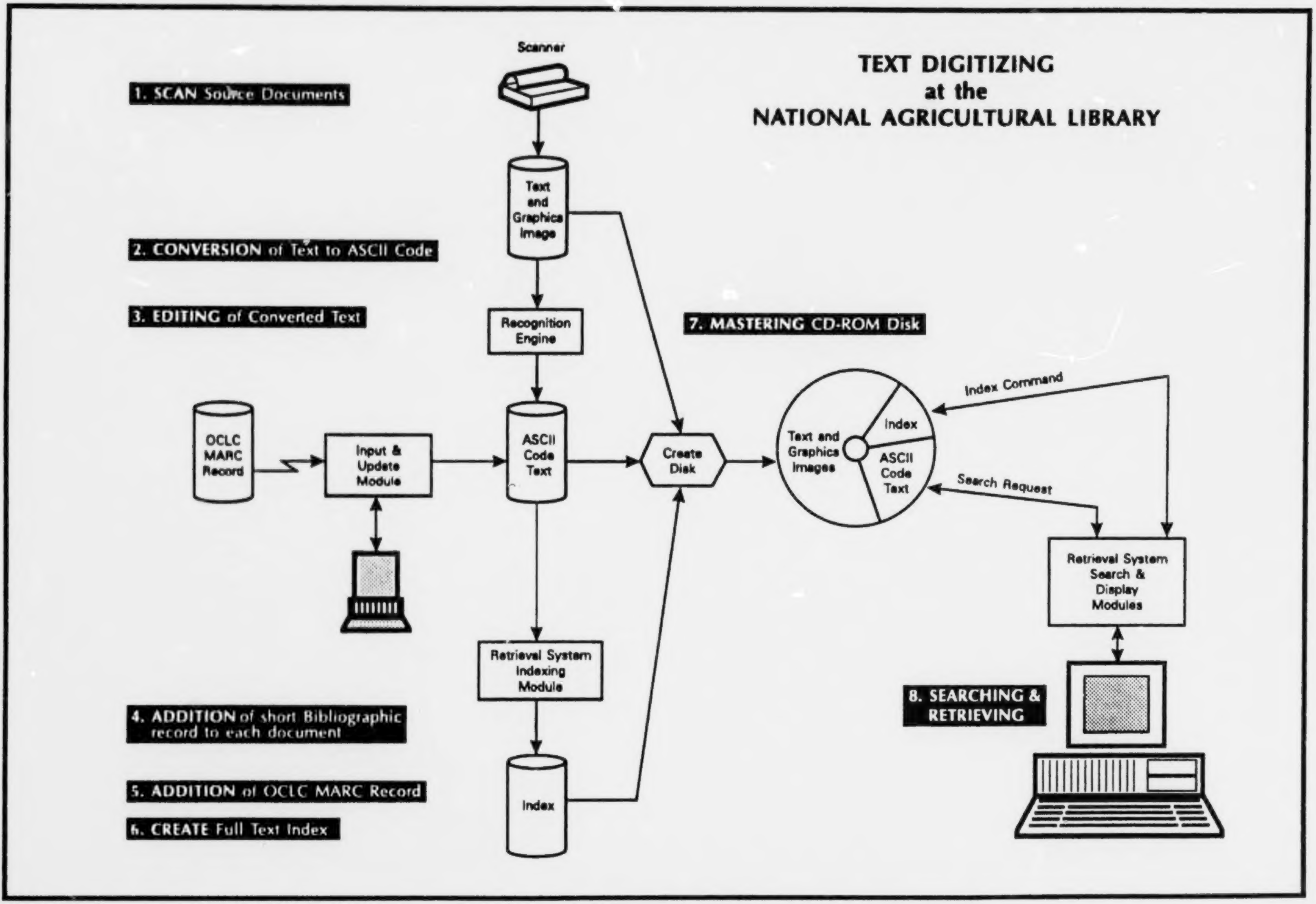

• Documents
• Drawings
• Mixed Text and Graphics
• Scan
• Store Images
• Convert Text to Machine-Readable Form
Verification Edit
Full Text Indexing
TEXT RETRIEVAL DATA BASE
IMAGE DATA BASE
WORM DISK 6,500 Pages of Image Plus Converted Fully Indexed Text (5 1/4 Inch Disks)
CD-ROM 6,500 Pages of Image Plus Converted Fully Indexed Text

Diagram 1

Diagram 2


Form Feed Flat Bed Scanner
✓ 3"x5" To 11"x17" Documents
✓ 300 DPI
✓ Paper Handler up to 50 Sheets
✓ 2.6 Second Scan Rate for 8.5"x11"
PC
✓ Hi-Resolution Graphics/ Text Monitor
✓ Display Image and Text Format
✓ Image Verification
Hard Disk Storage
✓ Hard Disk Storage
✓ File Server
CDP Controller
Optical Disk Drive and Controller (Worm)
Recognition Server
✓ Convert Image Text To Machine Readable Text (ASCII)
Information Retrieval Software
✓ Full Text Search
✓ Display & Print image and/or Text Formats



TEXT DIGITIZING at the NATIONAL AGRICULTURAL LIBRARY

Scanner
1. SCAN Source Documents
Text and Graphics Image
2. CONVERSION of Text to ASCII Code
3. EDITING of Converted Text
Recognition Engine
OCLC MARC Record
Input & Update Module
ASCII Code Text
Retrieval System Indexing Module
Index
4. ADDITION of short Bibliographic record to each document
5. ADDITION of OCLC MARC Record
6. CREATE Full Text Index
7. MASTERING CD-ROM Disk
Create Disk
Text and Graphics Images
Index
ASCII Code Text
Index Command
Search Request
Retrieval System Search & Display Modules
8. SEARCHING & RETRIEVING

Diagram 3

40Mb hard disk, 1.2Mb floppy drive

- LaserView 17" High-Resolution Portrait Monitor
- LaserData Compression/Decompression Processor
- Ricoh 4081 Laser Printer
- Maxtor 5-1/4" WORM Drive with caching disk
- Sony CD-ROM Player CDU 100B (added in January 1989)

Software

- MS-DOS 3.2 Operating System
- LaserView Image Handling Software from LaserData
- LaserBank WORM Management Software from Atlantis Research
- MS-DOS Extensions 1.0 (added in January 1989)

The initial cost for these two workstations and peripherals, including the customized SAIC operations software and SAIC support activities, was about $200,000. See Foreword by Clifford A. Lynch for more comments on this.

Upgrades to the workstations were made at various times throughout the pilot study, primarily in the form of software. In March 1988, WordPerfect 4.2 was integrated into the operations software for text editing, since the editor which originally came with the software was inadequate. In May, the two workstations at NAL were networked together, using DNA software, to facilitate the transfer of files for editing. The next major upgrade was performed in June 1988, when the Calera RS 3000 scanner was upgraded to an RS 9000. The RS 9000 software was faster and more accurate for text recognition, reducing the text editing load somewhat. During this period, additional modules of the operations software were installed, such as the software for archiving documents to the WORM and the indexing software, but these were not considered upgrades as such.

Until June 1988, upgrades were made primarily to reduce the time-consuming and labor-intensive task of text editing. In July 1988, after discussions between NAL and SAIC, major changes that affected the entire processing routine were made to the menu-driven operations software. The new menu system was not compatible with the old, so documents scanned under the old software had to be processed to completion (indexed) before the new version of the software could be implemented. The new software had several benefits:

1. It could track more than one data set at a time. This allowed the NATDP staff to work on two or three projects at a time, with little danger of documents from different databases being accidentally mixed together into the same database.

2. It archived the image data to WORM very early in the process, instead of at the end, leaving more space free on the hard disk.

3. It automatically updated the ASCII text file when changes were made to the corresponding image files, saving a great deal of time during text editing.

4. It recorded in the ASCII text at the bottom of each page the file name of the image file which corresponded to that page. This had great potential benefit, and was used in the *Agent Orange* database a year later to link certain images to the text files.

The new software had one major design flaw: The archiving of images to WORM could no longer be done as batch jobs of twenty or thirty documents, but instead had to be done one document at a time. Because each document contains many image files and image files are relatively large while WORM drives are slow, this added many hours to the processing time. (Archiving had formerly been done overnight, while staff members were off duty.)

The operations software was upgraded two more times, primarily to take care of serious bugs, but the menus remained the same. So, while these upgrades required some further testing, they did not require further training.

Another area that required upgrade was the WORM subsystem. The WORM software, LaserBank Link, proved to be unreliable. Occasionally, large files written to the WORM contained zero bytes; other times, the files disappeared altogether. This was a serious problem, since the system used the WORM optical disk as the only permanent data storage device. In January 1990, LaserBank Link was replaced by a Corel interface. Although this necessitated copying the existing WORMs to the new format, the result was a reliable archival subsystem. After writing thousands of files to WORM, the Corel software has not made a single error, and the interface is easier to use.

The hard disk and various boards were replaced in February 1989 after two serious hard disk crashes. Although these were more repairs than upgrades, the board replacements required reintegration of the entire scanning system. The "new" system proved to be more reliable, and has not suffered a breakdown for over twenty months—a remarkable record considering the system's complexity. This may be due partially to the fact that it was moved to a cooler area of the Library (as recommended by SAIC), and a fan was placed to blow continuously on the CPU. In addition, an Uninterruptable Power Supply unit was installed to protect the 384Mb hard disk against power surges and failures.

Finally, the 9-track tape drive unit proved to be inadequate for disk premastering. Its software could write tapes only at a density of 1600 bpi, which was much lower than the 6250 bpi needed for NATDP's large databases. Rather than upgrade the tape drive unit, NATDP elected to use a CD-Publisher purchased by NAL for use on other projects. This proved to be a good approach, since the CD-Publisher's high-density tapes could be used by all the CD-ROM mastering facilities.

Vendor support on the project was provided by the system integrator, SAIC. It was extensive in the first year, but the necessity for it tapered off as the NATDP staff became more knowledgeable and the system more reliable. In the first year, support included setting up the workstations, providing maintenance service, supplying written documentation for system operations, and training the NATDP staff. In addition, vendor staff assisted with image scanning, text editing, and indexing for the first CD-ROM, *Aquaculture I*. This "hand holding" type of support was not needed after the first year, but system maintenance support has continued throughout the pilot study.

Support was also provided by other vendors during production of the first disk. Meridian Data (supplier of the CD-Publisher used to premaster the CD-ROMs) and Digital Audio Disc Corporation (the CD-ROM mastering facility used for *Aquaculture I*) provided advice by phone concerning the disk premastering process, disk formats, tape formats, etc., that greatly facilitated production of the first disk.

Scanning Experience

The Central Scanning Facility is managed by one full-time NAL employee, the Project Coordinator. An additional employee serves part time as the System Manager. The center has been staffed throughout the pilot study by students from the University of Maryland. There are usually two part-time students, but on occasion, there have been as many as five. Because of the temporary nature of their situation, the turnover rate has been fairly high. Each new student must be trained, and this can tax the rest of the staff. During the first year, one of the employed students was from Maryland's School of Library and Information Science, and served as a Graduate Assistant on the project. Her contribution to the success of NATDP was significant. All the students have been enthusiastic, hard working, and dedicated. They have suggested improvements in the work flow and processing of materials, and their energy and interest have been assets.

The goal of the Central Scanning Facility was to create an electronic database from a collection of printed materials and distribute that database on CD-ROM. The process was as follows:

1. Collect appropriate publications.
   (This was usually done by a reference specialist in NAL's Public Services Division.)

2. Design the database.

3. Prepare publications for scanning.
   (This included preparation of worksheets for each record or publication. In some cases, subject descriptors were added either by the students or by a specialist in NAL's Indexing Branch.)

4. Scan page images and input relational data from worksheets.

5. Perform text recognition to create ASCII text files.

6. Verify ASCII text files and archive images to WORM.
   (Once the operator verified that all the appropriate pages in a document had been converted to text, the system automatically archived the document's page images to WORM.)

7. Review and edit ASCII text.

8. Mark ASCII text file as "Complete".
   (Once this was done, the system automatically attached to the ASCII file the relational data that was keyed in at time of page scanning.)

9. Archive ASCII text files to WORM.

10. Index the ASCII text files so they can be searched.

11. Review and test the database.

12. Prepare premaster tapes.

13. Master the CD-ROM.
    (This was performed by a commercial mastering facility.)

14. Prepare database documentation.

15. Distribute CD-ROMs with appropriate documentation.

The original project plan did not include text editing. The industry-reported error rate for text recognition was about one percent. Our databases would include both text and page images, with the text being used for search and retrieval, and the images for display and printout. A 99 percent accuracy rate was considered to be adequate for retrieval purposes, and the page images would provide information to the user at 100 percent accuracy.

Error rates for the materials NATDP was scanning were from one to 10 percent, with a few very poor publications running 15 to 20 percent, rather different from those advertised by vendors. One single-page publication that was of particularly poor quality had an error rate of 71 percent. These error rates were thought to be too high for accurate retrieval. The next section of this report contains a discussion of an error rate study performed by NAL staff.

In addition, the SAIC system provided an image format system which was proprietary, and we only had software which could be displayed on the high-resolution (150 dpi) LaserView monitor we were using. This meant that sites which did not have the LaserView monitor could not display the page images. Only five of the forty-six evaluation sites had the LaserView. Therefore, most of the sites would have to depend on the ASCII text files for both retrieval and display.

For these two reasons, it was decided that the staff should edit the ASCII text files. Text editing was done in WordPerfect 4.2. It was the most time-consuming of all the data preparation tasks, taking about three times as long as the scanning and text recognition tasks combined. In addition to correcting character recognition errors (such as "ii" substituted for "m", etc.), the editors had to delete "noise" characters, which we commonly referred to as "graffiti". Graffiti consisted of nonsense characters generated in the ASCII file when a page image contained a picture or chart embedded in the text, the shadow from a book binding, or other non-textual artifacts. Occasionally, page layout errors occurred in ASCII files generated from page images with complex, multi-columned layouts. Text from different columns might be mixed together, or decolumned text might have paragraphs arranged out of proper sequence. For information such as equations or chemical formulae which could not be properly represented in ASCII, a message to "See Page Image" had to be inserted. A similar message was inserted when an entire page was omitted from the text file (such as a page whose entire contents was a photograph). In these cases, specially developed macros were used to insert the substitute messages.

For the first CD-ROM, *Aquaculture I*, the staff experimented with three levels of text editing. Level 1 = Full editing. Level 2 = Partial editing (remove graffiti, correct page format errors, insert necessary messages). Level 3 = No editing (text used as is). The goal was to determine the degree of retrieval accuracy with text edited at the three levels, and to test user response to unedited or partially edited text. One surprising result was that the unedited text played havoc with the indexing software (TextWare Plus) used for the first disk. Text created for all subsequent disks was edited at Level 1.

Although the editors were proficient, it had been decided by the summer of 1989 that material with worse than a 10 percent error rate should not be converted to ASCII at all, because of the time and resources needed to edit it. Very old publications which were yellowed or brittle, with print bleeding into the page or poor contrast between the print and the paper, were among this material. Many of the older, deteriorating publications were the very documents that NATDP had envisioned for electronic capture and preservation. For this older material, the pages were scanned as page images, but text recognition was not performed. Instead, bibliographic records were keyed in manually or downloaded from NAL's bibliographic database for use in search and retrieval.

The costs for processing publications by NATDP, including administrative costs and overhead plus cost of equipment amortization, were as summarized in Table 3:

Table 3

Publication Processing Costs

Item		Cost/Page
Data preparation and capture, excluding text editing		$ 2.86
Editing ASCII text -	Level 1	5.32
	Level 2	2.54
	Level 3	1.54
Mastering CD-ROM with 100 copies ($3,000) at 6000 pages per disk		.50

Processing costs have gone up slightly over the life of the project, as salaries and supply prices have increased. But the cost to master a disk has gone down, with the cost now around $1,000.00 plus $2.00 per copy.

OCR Error Rate Study

One of the challenges to the project staff was to identify materials that could be successfully scanned and converted into ASCII form. To better identify such materials, an error rate study was performed to determine the characteristics associated with successful conversion. Paper types, print styles, and format parameters are among the factors studied. The study included two thousand pages of scanned and converted aquaculture materials. These comprised 156 documents which represented chapters, major headings, and small publications such as pamphlets. Documents were classified according to the following characteristics: contrast between ink and background; paper characteristics; quality of letters; print medium; type size and type families; characters per column; spacing; and format (see Table 4).

The spell check module of the word processing package WordPerfect was used to identify errors. Spelling errors were counted and compared against a word count corrected for the amount of graffiti. There are some limitations using this method to find errors. The spell check will not identify numeric errors, words requiring capitalization, and words correctly spelled but inserted in place of those included in the original document. Therefore, it was necessary to check all the numbers in the document.

Frequent count and various measures of association were made using a statistical package. Contingency tables displayed the joint frequency distribution according to two variables: error rates and specific document characteristics.

In terms of sample characteristics, documents selected for this study had similar characteristics, with the exception of one glossy paper. The majority (93 percent) had good contrast between the black ink and white paper. Most documents (65 percent) displayed a minimal amount of bleed-through. Typewritten documents (59 percent) outnumbered typeset documents (34 percent), and seven percent of the total documents were computer generated. Two percent of all the documents were photocopies.

In terms of type families, sans serif (54 percent) and modern (44 percent) type families were primarily used. Sans serif families are exemplified by Helvetica and Futura, modern families by Times Roman. There was not much difference in point types (range of 8-10). Most of the documents consisted of one column that was flush right (65 percent) and flush left (62 percent), with seventy characters per column (54 percent). Two point line spacing was the most frequently used spacing (45 percent). More than half of the documents consisted of text only (57 percent) and 28 percent had approximately 75 percent text.

To determine error rates, an analysis was made of misspelled words and numbers. The mean error rate for text conversion was 3.6 percent. The highest error rate for a single document was 71 percent. Upon reviewing the entire document, a number of significant relationships were noted. There is a significant relationship between the letter quality of the document and success of character recognition. Documents rated as poor or fuzzy had a greater frequency of errors. The poorer the letter quality became, the more the system misread the lower case "h" and to a lesser degree the letters "m" and "n". A higher error rate (greater than nine percent) was found for letter quality of the original than in the findings of a recent article (eight percent) on the performance of the Palantir Compound Document Processor (CDP).

The relationship between the number of errors and type families was also investigated. Documents using a sans serif face had a higher frequency of errors. As an example, spaces or graffiti inserted in words occurred more often in these documents than in documents using other type families. Sans serif lower case letters "a" and "o" were more often misread than a modern type for these letters.

Type of print also had some influence on the success of the character recognition. Each letter on a typewriter occupies the same amount of space. Typeset material, on the other hand, is spaced proportionally; i.e., the letter "l" will not take up as much space as an "m". An optical character recognition (OCR) device not set up for this kind of printing will have trouble finding the next character.

The type of formatting used also affected character recognition. A significant relationship was seen in documents with extra white space between paragraphs and the conversion of words with space and graffiti inserted in them. Documents with one point line spacing had fewer errors than those with two point spacing. The number of documents with one point spacing was evenly distributed among the three categories of error rate studied.

All eleven documents using a header to the right or left of the text had a spacing problem. The line containing the header and the text converted into one word. This appears to be a big problem.

Multi-columned documents also presented a challenge. Such variables as the number of columns, the inclusion of graphics within a column, and the presence or absence of significant spacing between columns can significantly impact the structure of the converted text.

Table 4

Characteristics Evaluated

Document

Contrast between ink and background
Paper quality
Paper finish
Bleed through
Quality of letters
Print medium: typewritten, computer generated, typeset
Document type
Photocopied

Print

Type sizes
Type family
Number of type families
Number of characters per column

Format

Spacing of letters
Flush right and left
Spacing of lines or leading
Number of columns
Amount of space between paragraphs
Underlining
Canceled type

In summary, what determines whether a document will be successfully converted with a high degree of accuracy? The partial answer to this question is letter quality, type of font, type of print, and format. There is a statistically significant relationship between letter quality of the document and the success of character recognition. Documents with sans serif had a higher frequency of errors, especially the occurrence of spaces or graffiti inserted in words. Sans serif letters "a" and "o" were more often misread than the modern family type for these letters. Type of print also had some influence on the success of the character recognition. Fewer errors were associated with typewritten documents. Formatting also had an effect. A significant relationship was seen in a situation when extra space occurred between paragraphs and the occurrence of words with spaces and graffiti inserted in them. Documents with one line spacing had fewer errors than those with two point spacing. Multi-columned texts occasionally presented a formatting challenge. Due to the similarity of documents, we did not have a large enough sample to test the significance of the following variables: dot matrix, headers to the right or left of the text, and signatures. Custom error-correcting OCR software would have been a fruitful avenue for future research, but NATDP was not in a position to pursue such research.

Scanning from Microform

As part of its ongoing investigation into preservation scanning, the NATDP tested the feasibility of converting microform to electronic page images. For this effort, a collection of microfilm containing the papers and letters of George Washington Carver was selected. The microfilm was optically scanned and converted to bit-mapped images and then distributed on CD-ROM.

Optical scanners that use microfilm as the source material came on the market around 1988. With special image-enhancing filters, knows as image optimizers, they can produce electronic images of the individual frames of a microfilm reel that are often superior to what is on the microfilm itself. The reel is mounted on the scanner much in the same way a reel of film is loaded into a camera. The operator makes certain adjustments based upon the characteristics of the microfilm (e.g., frame size, negative vs. positive image), and then starts the scanning process. The film feeds through automatically, and an electronic image is made of each frame at a speed of about three seconds per frame. The resulting images are then separated by the operator into logical groups; and each group will become a record in the image database.

NAL received a grant from USDA Science and Education Evaluation Funds to explore this process. Rather than purchase the expensive scanning equipment involved, NAL decided that the scanning itself should be done by NATDP's service contractor, Science Applications International Corporation (SAIC). Robert Butler, NAL Coordinator of Preservation and Access Programs, and Judith Zidar, Coordinator of NATDP, were assigned the roles of selecting the microfilm and coordinating the project.

In February and March of 1991, test scanning was done by SAIC of several reels of microfilm supplied by NAL. It was important to determine ahead of time which microfilm characteristics would affect image quality and whether there was a possibility of performing text recognition on the optimized images for purposes of indexing. Working with 35mm silver halide film, NAL staff evaluated negative versus positive images, landscape versus portrait orientation, and single versus dual-page frames. None of these factors significantly affected image quality, since the image optimizer was able to adjust for them. Even minor scratches and other imperfections could be corrected. Three factors that did impact the quality of the images were (1) the apparent quality of the original from which the microfilming was done; (2) the apparent care that was taken during the microfilming process; and (3) the amount of size reduction that was done on the contents of each frame. Some of the film resulted in images that were too poor for the human eye to read, even on a very high resolution monitor, although many of these also could not be read using a microfilm reader. None of the test microfilm reels produced images that, on the whole, were good enough for OCR treatment. Computer-Output-Microfilm (COM) might have given better results, but there was none available at the time of testing.

After completing the test scanning, NAL began the final selection process for material for the "microfilm CD-ROM". Since OCR was not to be used, other criteria predominated: the material must be informative and visually interesting, it must form a cohesive subject collection, and it must be valuable in its own right. The Tuskegee University microfilm collection of the papers and letters of George Washington Carver more than fulfilled these criteria. Carver is a major historical figure in agriculture who devoted himself to practical, hands-on agricultural research and education. On sixty-seven reels of microfilm, Tuskegee University (formerly Tuskegee Institute) had recorded Carver's papers, letters, records, and drawings of seeds, plants, and inventions. Much of his personality has been captured as well. With Tuskegee's permission, NAL placed the contents of three reels of these materials on CD-ROM, along with bibliographic information for access. In addition, the complete text of *Guide to the Microfilm Edition of the George Washington Carver Papers at Tuskegee Institute* were included on the disk and are fully searchable. The disk was distributed in December 1991.

Archival Format Standard

The original contract with SAIC required that all scanned text be stored in a standardized format so that it could be used independent of any workstation or any production system. Experience in this area was unsatisfactory, in that no national or international standard had yet been accepted for archiving images. A version of the Tagged Image File Format (TIFF) for images was used by SAIC in later stages of the project, but there were multiple versions of TIFF formats in use within the industry. In this area, the project did not fulfill its objective.

The Coalition for Networked Information, a group working toward developing the infrastructure for the National Research and Education Network (NREN) in the United States, is trying to establish a de facto standard using TIFF B format, which could be adapted by NAL as its standard. At present, this is an unresolved problem.

Workstations

The project distributed two versions of workstations: (1) a high-resolution workstation based upon the LaserData proprietary formats and the SAIC integration software, which produced a very high-quality graphic image and which could display a full-page image on the screen; and (2) a low-end standard personal computer which could display only ASCII text and which presented particular problems in how much text displayed, requiring difficult scrolling processes to see the full text. Because the high-end workstation used proprietary hardware and software which would be unaffordable for many users, and because the large base of equipment available to libraries was at the personal computer level, the project wished to test the acceptability of software and displays on a low-end, less costly workstation using off-the-shelf equipment and software.

High-Resolution Graphics Workstation

Costs for the initial workstation were $17,000.00. These costs included:

- NEC Powermate 80286 Microcomputer with 640Kb RAM, 40Mb hard disk and 1.2Mb floppy disk drives
- LaserView 17" High-Resolution Portrait Monitor
- LaserData Compression-Decompression Processor subsystem
- Ricoh 4081 Laser Printer
- Sony CDU-100 CD-ROM Drive
- MS-DOS 3.1 Operating System
- LaserView Image Processing Software from LaserData
- Browse & Print Image Handling Software from SAIC
- MS-DOS CD-ROM Extensions

The original high-resolution graphics workstations were upgraded in 1989 to include NEC MultiSync 2D monitors, Video Seven VGA display adapters with 256Kb memory, and Hewlett Packard LaserJet II printers (with 2Mb additional memory). This upgrade allowed viewing and printing TIFF images included on some disks. In 1990, Everex memory expansion boards with 2Mb memory and Logitech serial mice were added for the Windows-based *Agent Orange* disk. Total additional cost: approximately $2,000/workstation. Delivery of the original high-resolution workstations was made by SAIC. Workstations had software installed before shipping in all cases but one, thereby simplifying troubleshooting by NAL staff in the early phase of the project. One site received a high-resolution workstation late in the project. This system had no software installed and system setup proved complex enough that a visit to the site was made by an NAL staff member to expedite the setup. System hardware upgrades caused integration problems at the sites, requiring extensive telephone support by NAL staff. Problems were usually solved rapidly but some difficulties lingered, requiring several calls before a solution was reached.

Workstation strengths and weaknesses are both due to the same factor: the LaserData high-resolution display subsystem. The advantages inherent in the excellent screen and print resolution available through the LaserView monitor and proprietary image format come at a high cost in DOS memory and expensive microcomputer hardware. These advantages include the portrait-oriented monitor, full-page display, and paper-white image, all of which are expensive hardware solutions as implemented by LaserData. LaserData's proprietary image format permits very tight compression allowing low mass storage requirements. Image retrieval and processing is rapid due to the use of hardware for compression and decompression as opposed to software-based solutions which are generally slower. After loading all the necessary device drivers and terminate-stay-resident (TSR) software to use the high-resolution monitor and printer and Microsoft CD-ROM Extensions, less than 500Kb memory remains available to DOS. Adding the memory expansion board device driver uses more memory, making the installation and running of several software packages challenging. One search and retrieval package required a minimum of 517Kb free DOS memory. NAL staff became adept at reconfiguring the workstations to reflect the different setup requirements for each retrieval package. SAIC provided a smaller version of the TSR software for image handling that frees some of this DOS memory, but these memory problems are still serious. Our original workstations were NEC 80286 class machines running at 10mHz clock speed. Since these are not the cheapest or latest development in workstations, NAL staff began investigating other microcomputer platforms. LaserData informed NAL staff that the only microcomputer approved for the LaserView subsystem was the NEC. NAL staff had no success in integrating the imaging subsystem into a COMPAQ 386/20 and did not try to install this subsystem into a 386 clone. NAL is researching alternative imaging systems that may ameliorate some of these problems and extravagant hardware and memory requirements. Recent developments regarding CCITT Group 4 compression of TIFF images may prove to be the answer to this image format problem, allowing the display of page images on any graphics monitor.

ASCII Workstation

Initially, sites with ASCII-only workstations received a SONY CD-ROM drive and Microsoft CD-ROM Extensions for that drive. This was a hardware/software component that challenged the ingenuity of staff at several of the sites as the integration of the drive proved difficult on their microcomputer hardware. NAL staff assistance to sites was considerable, primarily by phone but also in written documentation and correspondence. The majority of the problems reported by sites were concerned with successful installation of the MS CD-ROM Extensions delivered with the Sony CD-ROM drive as the default settings were not correct for the DOS microcomputers not manufactured by SONY. Many hours were spent trying to decipher the cryptic instructions delivered with the SONY drives. One project coordinator traveled to NAL with his CD-ROM drive and microcomputer for troubleshooting; the problem proved to be a bent pin on the CD-ROM drive cable. Several of the reported problems were traced to this type of hardware failure or setup problem. Sites' hardware included microcomputers from IBM, COMPAQ, Leading Edge, Zenith, and other manufacturers. System architecture has included 8088-, 80286-, and 80386-based systems, with the expected improvements in performance on higher level machines.

Upgrades provided to the ASCII-only sites by NAL have been limited to software improvements and suggestions (e.g. adding memory to laser printers increases printer performance with some search software packages) mostly by phone, documentation, and correspondence. Some sites have opted to upgrade hardware with expanded memory for their microcomputers and/or more memory for laser printers to view and print TIFF images included on some of the disks. Strengths of the ASCII-only workstations include low price and the ability to use currently owned hardware. Limitations of the ASCII-only sites include inability to view high-resolution graphics and a not-inconsequential level of difficulty with integration due to different levels of "IBM compatibility." During the early part of the project, the microcomputer procurement policies of some of the libraries created rather interesting situations for the site coordinator and NAL staff when trying to troubleshoot problems. One site using a Zenith microcomputer was never able to install the Sony drive successfully; they finally used a different type of microcomputer. The different types of hardware used by the ASCII-only sites made NAL assistance appear unequal due to lack of experience with different "IBM compatible" platforms; since most sites have available expert help on campus for microcomputer hardware difficulties these local experts have proven useful to NAL staff when integration problems arise. With the introduction of a Windows 3.0-based application, many sites have complained about the need to upgrade their hardware to 80386-based high-RAM microcomputers; this seems to reflect a rather narrow viewpoint in light of the current low price of 80386 microcomputers and the rapid advancements in the field of microcomputer design. The future products from the project will probably have a text-only version of the search software for those sites not willing to upgrade their hardware.

Workstation White Paper

Because of the compatibility problems with the workstations, Dr. Clifford Lynch was commissioned to prepare a white paper on workstation platform issues to be examined as NAL considered the environment in which this technology would gain wide acceptance. The original white paper is attached as Appendix 4 and is updated in Dr. Lynch's Foreword to this final report.

Software Package Evaluation

When this project began, there was very little software available that could handle bit-mapped images or could cross-index between ASCII text and its corresponding bit-mapped page image. The first package used, TextWare, was one of the few commercial packages available in the marketplace in 1988. The packages chosen for each test disk reflect what was available in the market at the time. Those options changed dramatically during the course of the field tests.

At the end of the test period, so many new packages were available that the project commissioned an independent study which evaluated thirteen software packages for future use by NAL for production disks. That study was conducted by Pauline A. Zoellick, who reviewed the thirteen software packages and eventually recommended three as coming closest to meeting the specifications developed for the evaluation. A panel of five librarians from test sites was called together to aid in the comparisons. The NATDP Software Evaluation Team met at NAL in Beltsville, Maryland during October 15-19, 1990 and tested all three of these products after receiving vendor business and technical presentations. The panel then ranked the three software packages. Based upon that process, a licensing agreement has been negotiated with Personal Library Software, Inc. for production of future disks. Section I (Statement of Work and Work Products, and Methodology) and Section II (Factors for Consideration in Selecting a Vendor) of the consulting report are included as Appendix 5. Analysis of the specific vendors has been omitted from this report due to the proprietary nature of the information and due to the fact that such information dates very quickly.

Field Experience with Demonstration Disks

A major objective of the project was to test this technology with real users in the field—librarians, library patrons, and researchers. There was also interest in working with the World Bank in testing CD-ROM technology in international research stations where telecommunications were limited, to see whether researchers under those conditions would show a higher tolerance for some of the limitations of the technology.

There was no expectation that the field tests would provide a statistically valid sample, as library sites were able to provide only a small test base willing to spend the time required to use the test scripts and provide results. However, the test base was large enough to provide extremely helpful objective evidence of user acceptance.

What became more obvious during the test period was that the tests provided a necessary learning experience for libraries in the requirements to mount such projects and to support them technically. This was also true during Phase 3 with the Internet telecommunications project.

The field tests were conducted using formal scripts and separate questionnaires for installers, librarians, and end-users. Results of the field questionnaires are summarized below.

Aquaculture

The first collection scanned, digitized, and distributed to participating libraries for field testing was a collection of 4,000 pages of non-copyrighted aquaculture materials, with TextWare as the search and retrieval software.

To evaluate the quality of the database and its retrieval software, local project coordinators, librarians, and various groups of end-users at participating institutions were recruited to search the database and to complete an evaluation questionnaire. Generally speaking, the quality of the optical character recognition for this collection was found to be less than acceptable, with a significant percentage of the database having errors in conversion from printed source material to ASCII format. In a comparison of all systems evaluated in the field tests done in the *Agent Orange* evaluation, a majority of the respondents rated this database and software "unacceptable." Some of the problems included difficulty in moving around in the software, inadequate documentation, and inadequate screen guides. Users often would get lost and be unable to get out of certain functions.

Food, Agriculture, and Science

In 1989, a second disk, *Food, Agriculture and Science*, was prepared in cooperation with the Consultative Group on International Agricultural Research (CGIAR). This collection, a sample of publications from selected international agricultural research centers, was distributed in the fall to forty-fix sites, with KAware as the search and retrieval software. By the spring of 1990, a total of 162 evaluation questionnaires were received from twenty-eight sites. Of the questionnaires received, sixty-nine were completed by librarians, sixty-eight by end-users (including twenty faculty, nineteen graduate students and twelve undergraduates), and twenty-five by the local project coordinator or other staff designated to install the system.

Most of the coordinators (or other designated staff) had no problem with the installation of the KAware software. Nearly half (48 percent) found it "very easy" to install, while one-third (35 percent) found it "somewhat easy." Twenty percent of the coordinators were able to install the software in fifteen minutes or less, while 65 percent were able to do so in thirty minutes or less.

During the installation and testing of the software, a significant percentage (17 percent) of the coordinators encountered problems with the installation, or could intentionally make the system 'crash' (54 percent). Several coordinators commented on interference with other software on the same computer or insufficient memory for execution.

While KAware was seen as an improvement over the *Aquaculture* disk and its search and retrieval software, the file organization of the database, the KAware software, and selected system features were found to be less than acceptable by a significant percentage of librarians and end-users. Overall, only 53 percent of the librarians found the system "very easy" (four percent) or "somewhat easy" (49 percent) to use. End-user responses were comparable: 59 percent found the system "very easy" to use (10 percent) or "somewhat easy" to use (49 percent). Of those responding, only 65 percent of the librarians found the function keys used to manipulate the database to be understandable. Only slightly more than half (57 percent) of the librarians responding found the system screens acceptable, while an equal percentage of those responding ranked the software overall as "easy to understand." End-user results were mixed: 59 percent of those responding found the screen layout acceptable, while 85 percent found it easy to read. Nearly forty percent (38 percent) of local project coordinators, however, found screens to be "cluttered" (13 percent) or "somewhat cluttered" (25 percent).

For the 'Table of Contents' function, a feature that allows users to scan through a listing of publication titles and their table of contents contained within the database, only two-thirds (66 percent) of the librarians responding found this feature acceptable. About half (48 percent) of this group, found it difficult to 'move' within this feature to locate an entry for a specific 'book' title requested in the evaluation's guided tour, a modified tutorial that instructed users how to use specific features and to perform specific searches. End-users were less perplexed: 74 percent of those responding found this feature acceptable and 69 percent found it "easy" or "somewhat easy" to move through the listing of titles. Interestingly, end-users generally tended to be less critical than librarians of this and other features evaluated in the field tests.

Image management generally received high ratings from both librarians and end-users. Nearly 83 percent of the librarians responding found the image management features acceptable, while 95 percent of the end-users responding found it easy to call up a bit-mapped graphic page image.

Some of the general perceptions noted by the respondents included the fact that only about 50 percent of the librarians felt it was apparent where they were when they completed a search. Other problems noted by the two user groups included: difficulty distinguishing between records, cluttered screens, poor screen prompts, 'Help' screens which did not relate directly to the screen or the problem encountered, poor instructions and tour, a variety of print problems, and the lack of a proximity search.

Acid Rain

In the spring of 1990, a two-disk set containing a collection of several hundred Canadian government documents on acid rain, with an enhanced version of the search software, KAware2, was sent to Iowa State University for the preparation of an appropriate guided tour and questionnaire. The transfer of the project from the University of Vermont and the response from the initial field tests from the *Aquaculture* disk offered an opportunity to revise the evaluation materials for the remaining field tests. Diana Shonrock, Bibliographic Instruction Librarian at Iowa State University (ISU), who had considerable experience with writing evaluation instruments and with statistical methods, was able to incorporate more of the user perspective in the overall design and content of the tour. Revisions of the tour and evaluation form were completed with assistance from Gerard McKiernan of the ISU Reference Department and from the personnel of the Statistical Lab at ISU including Roy Hickman and Kathy Shelley.

To permit comparison of data from each of the field tests, it was decided that the guided tour for the *Acid Rain* disks would be as similar in structure and, where possible, in content, to the tour prepared for the *Food, Agriculture and Science* database, the CGIAR disk. To compare responses across user groups, it was also decided to use the identical evaluation questionnaire for librarians and end-users; in the previous evaluation, different questionnaires were used for different user groups. To determine the relative acceptance of search and retrieval software used in this and subsequent evaluations, sites were also asked to recruit as many of the same users for each of the field tests as possible.

A special effort was also made to incorporate recommendations made by users from the previous tests, as well as to make improvements based on user reaction to the guided tour and the system during the field test of the CGIAR disk at ISU. Casual observations of user behavior, as well as informal comments made during this local test, brought to light inadequacies in the guided tour. A number of these problems confused the user and no doubt affected user acceptance of the search software and the system in general. Nationally, in the CGIAR test, 38 percent of the librarians had found the tour difficult to use. These subjective expressions were corroborated by local project coordinators who either "often" (17 percent) or "sometimes" (71 percent) found the documentation prepared for the CGIAR system to be confusing.

Tables 5-6 summarize responses by users of the *Acid Rain* disks. The number of questionnaires returned for the *Acid Rain* disks were comparable to those returned for the *Food, Agriculture and Science* database. For this evaluation, eighty-one questionnaires were completed by librarians, fifty-nine by end-users, and twenty-four by project coordinators or their designees.

The ease of installation for this search software was more than 10 percent higher than the rating in the previous field test: 96 percent of coordinators found this software "very easy" or "somewhat easy" to understand. About half (52 percent) of the local project coordinators were able to install the software in fifteen minutes or less, while 70 percent were able to do so in thirty minutes or less. Interestingly, however, about one-third (30 percent) of the coordinators encountered a problem during the installation process, nearly double the percentage reported with the previous disk.

An analysis of key responses from this evaluation demonstrates an overall increase in acceptability of the KAware2 software and this database.

Table 5

Characteristics of the *Acid Rain* Users

	End-Users	Librarians
Had searched CD-ROM before	78%	98%
Had searched full text before	60%	83%
Had used the CGIAR disk	20%	50%

Table 6

Degree of Expertise with the KAware Software

	End-Users		**Librarians**	
	CGIAR	**Acid Rain**	**CGIAR**	**Acid Rain**
Expert	na	4%	--	1%
Intermediate	na	47%	35%	56%
Novice	na	49%	65%	43%

Nearly four-fifths (79 percent) of the librarians found it "very easy" (14 percent) or "somewhat easy" (65 percent) to learn the basic features of the software, perhaps affected by previous use of the CGIAR system. End-user acceptability was even higher, with 90 percent finding it "very easy" (29 percent) or "somewhat easy" (61 percent) to learn the system. Also notably higher was the overall satisfaction with the guided tour; 77 percent of the end-users and 73 percent of the librarians felt it was adequate, compared with 59 percent and 62 percent of the respective groups for CGIAR.

Of those responding, nearly ninety percent (89 percent) of the librarians found the function keys understandable and 83 percent found the screens easy to read. End-user responses were comparable: 95 percent found use of the function keys "clear" (59 percent) or "somewhat clear" (36 percent), while 90 percent found the screens "very easy" (56 percent) or "somewhat easy" (34 percent) to read. Overall, user acceptance of the 'Table of Contents' feature was uniformly higher than its use with the CGIAR database: 85 percent as compared to 66 percent for the librarians and 88 percent as compared to 74 percent for the end-users found this feature acceptable.

Over 91 percent of the librarians and 83 percent of the end-users found it easy to perform a sample 'Subject' search using an appropriate Library of Congress Subject Heading. For a 'Proximity' search, a search in which users were asked to search the *Acid Rain* database for two terms occurring within a specified number of words of each other, 83 percent of the librarians and 91 percent of the end-users found it easy to perform this search. User acceptance of a sophisticated search feature, the 'Hypertext' function, was equally as high: 88 percent of the librarians and 93 percent of the end-users found it useful to 'jump to' other occurrences of a term (or sets of terms) elsewhere in the database to find additional related text. In contrast, in the field test for the CGIAR disk, librarian acceptance was comparable (89 percent) to that for the *Acid Rain* software, but end-user acceptance was more than ten percent less (81 percent).

The ability to manipulate the features of the image management system was comparable or higher than exhibited with the *Food, Agriculture and Science* database. Ninety-two percent of the librarians and 96 percent of the end-users found it easy to display an image, while nearly 94 percent of the librarians and 92 percent of the end-users found it easy to magnify an image to a legible size.

While there were marked increases in user acceptability of many features using an improved version of the search software and an enhanced guided tour, there were a number of important functions and features that received lower levels of acceptability. For example, only 61 percent of the librarians and 71 percent of the end-users found the system response time acceptable. Such comparatively low acceptance can be partially attributed to the platform used at the test sites; most used a 286-level computer, the standard established at the beginning of the project. As a result, it is not surprising to find that librarians as well as end-users, rated the response time of a free-text, global search at comparatively low levels: 63 percent and 71 percent, respectively. This low response might also be attributed to the inherent technical limitations of the software in searching for variant forms of a given word.

Slow processing speed no doubt also contributed to a comparable lower rating for overall acceptability of image management by both librarians and end-users. While users clearly found it easy to display an image, only 75 percent of the librarians and 81 percent of the end-users found the image management system acceptable, a decrease of 15 percent from the image manipulation ratings. This lower level of acceptance may be attributed to the fact that in KAware2, bit-mapped page images are not identical in size or dimension to the corresponding original document. To read easily a portion of text or view a detail from a page, users are required to magnify selected sections of the graphic image.

A similar limitation was identified by users in examining portions of the ASCII text displayed after a specified search of a document's full text. Several librarians and end-users considered the screen display cluttered or felt it was an unnecessary burden to have to manipulate the screen features to view the entire text of a particular document. Nearly half (46 percent) of the local coordinators found the screens "unclear." Clearly a better platform—a faster computer (386-based or higher) and a larger screen monitor—would help reduce these inadequacies.

In general, the users' favorite features of KAware2 included the ability to project images, the ability to mark sets, hypertext, and the 'note' function. The lack of speed, image quality, cluttered screens, lack of pure Boolean operators, and "Help" screens which were too general to provide the assistance needed were typical inadequacies mentioned. Overall, the librarians tended to be less accepting of the various features than the end-users, possibly because they had more experience with the other database retrieval systems. Librarians, in their comments, also indicated that they felt that the system would be too difficult for most end-users, although end-users themselves did not indicate such difficulty. "I am impressed by the sophistication of the software, but I frankly think it will blow away even the most sophisticated of end-users" and "For users without a step-by-step guide card, it would be difficult to understand" are examples of librarians' comments.

Agent Orange

The last disk to be evaluated was a database of special studies, reports, and similar documents, as well as selected copyrighted articles, relating to the defoliant Agent Orange. Windows Personal Librarian, a Windows-based text and image management system developed by Personal Library Software, Inc. (PLS), was the retrieval software used in the final field test. Only about one-fourth of the cooperating institutions (twelve of forty-six) participated in this evaluation; twelve coordinators, forty-six librarians and twenty-two end-users participated (less than half the number of those that participated in the field test of the CGIAR disk and the *Acid*

Rain disks). This relatively low level of participation may in part be attributed to the use of Windows, an operating environment that was not commonplace in libraries at the time of the field test, the level of hardware required, the length of the tour, and the amount of time required. The comparative sophistication and complexity of the search software itself also contributed to the low level of participation; one-third of the sites (four of twelve) found it difficult to install the software, while two-thirds (eight of twelve) encountered problems with the search software and/or the Windows software (four of twelve).

Tables 7-9 summarize some of the user data from the *Agent Orange* field tests. Of those completing an evaluation questionnaire for this system, 93 percent of the librarians and 86 percent of the end-users felt they were able to learn the basic features of this system, compared to 85 percent of the librarians and 88 percent of the end-users who were able to learn the basic features of the CGIAR system and 96 percent of the librarians and 97 percent of the end-users who were able to do so with KAware2 software used with the *Acid Rain* disks. Seventy-two percent of the librarians and 75 percent of the end-users found the *Agent Orange* tour adequate.

Overall understanding of this system was also lower; only two-thirds of all users (64 percent of the librarians and 68 percent of the end-users) found the system easy to understand. It is possible that general understanding and acceptability of the system and its features and functions is related to a user's familiarity with a feature, previous exposure to the Windows operating system, an understanding of the multi-tasking environment, and participation in previous field tests of the other collections or use of online and/or full text databases. For this field test, 85 percent of the librarians and end-users rated their searching ability as either "very proficient" (33 percent) or "somewhat proficient" (52 percent).

A major benefit of the PLS software is that it supports multi-platform systems (DOS, UNIX, Macintosh). Utilizing high-end hardware to address this problem during the research and development project was deemed acceptable, as the rest of the installed base would follow in the next few years.

Table 7

Characteristics of the *Agent Orange* Users

	End-Users	**Librarians**
Had searched CD-ROM before	86%	100%
Had searched full text before	67%	84%
Had searched the CGIAR disk	28%	55%
Had searched the *Acid Rain* disks	33%	70%

Table 8

Degree of Expertise with the *Agent Orange* Software

	End-Users	**Librarians**
Expert	---	---
Intermediate	41%	44%
Novice	59%	56%

It is not surprising to find that 81 percent of the librarians and 85 percent of the end-users found it easy to manipulate the menus during the *Agent Orange* tour; menus were used in the previous software and had become a feature common to many other software packages. To a lesser degree, the same may apply to acceptability of the 'Proximity' feature; while somewhat different than those used in the previous search software, and somewhat cryptic in its use of indicators and syntax, 77 percent of the librarians and 89 percent of the end-users found this search easy. Comments, however, indicate a level of dissatisfaction with the way results are displayed when the system highlights all words searched in a proximity search regardless of the actual proximity of words to one another.

When it came to unfamiliar features or functions in the Windows environment, users were not as satisfied. For example, the 'Lights' window, a control feature, was found to be unclear by 39 percent of the librarians and 42 percent of end-users. A comparable percentage of librarians and end-users found the bar charts that resulted from the algorithm 'Ranking' feature useful (69 percent and 63 percent, respectively) or clear (67 percent and 61 percent, respectively).

The 'Dictionary' feature and associated 'Tag' and 'Include' functions in the Windows Personal Librarian software are examples of software features that received high acceptance, even though such features were not fully represented in the previous evaluation and are not common in other software packages. This new feature, which allows the user to identify and select terms from an alphabetical listing of free-text terms for inclusion in a search strategy, was ranked high by users: 82 percent of the librarians and 90 percent of the end-users found the feature useful. Likewise, a task that required special use of the mouse—the 'Clearing' routine—was considered easy to perform by 79 percent of the librarians and 80 percent of the end-users.

Once again, however, response time was considered inadequate by a significant percentage of users. Over fifty percent (56 percent) of the librarians and a third of the end-users (33 percent) considered the system slow. This perception was corroborated by nearly identical ratings for the response time for multi-word processing. As with the previous field tests, this level of responsiveness can be directly attributed to the adequacy of the computer established as the standard for the project and the size and dimensions of the display monitor. Screen displays were also less acceptable than some features: only 74 percent of the librarians and 79 percent of end-users found the screens easy to read.

User reaction to this software generally reflected reaction to selected features of each of the previous systems as well as the unique features of a Windows-based operating environment and to the inability of the 286 hardware to react to commands at a speed which they felt was acceptable. One comment sums it up fairly well, "Waiting around for things to happen would try anyone's patience." Relating to the choice of the Windows environment, several users felt that an "expert" or function key driven option should be available. One users comment summed it up like this, "Pointing to pictures is cute, but typing is faster."

Several respondents had problems with the instructions/tour, many of them relating to the length and the necessity to explain new features. One suggestion which might improve this is the possibility of separating explanations from the actual instructions. The algorithmic ranking function was noted as a plus by several users, but a number of them indicated a lack of understanding of the concept behind the function. As ranking of documents is considered by many authorities a critical factor in identifying and selecting relevant documents or sections in full text databases, the algorithmic function should be made clearer and easier to use and promoted among users of full text databases. Other comments included some which had been common in earlier test systems: inability to feel comfortable with where they were in the database, slowness of the system, need to try to adapt the software and command language to an already familiar platform such as SilverPlatter or Wilson, and the cluttered appearance of the screens. Overall acceptability for *Agent Orange* software for full text data searching and retrieval broke down as follows:

Table 9

***Agent Orange* Software Acceptability**

	End-Users	**Librarians**
Very acceptable	28%	3%
Acceptable	56%	59%
Somewhat acceptable	17%	30%
Unacceptable	0%	9%

Agent Orange generally received positive response to the features which were unique to the Window's environment; however, a desire was expressed by those unfamiliar with it to follow a tutorial of the system before trying to use it for full text searching. Some users also commented on the need to "clean-up" the screens so they would be less cluttered and more representative of the original page. Because of the small number of evaluations returned, eighty of a possible 506, it is hard to determine the statistical significance of the data. Data reported are trends noted in the responses of the completed questionnaires.

Conclusions and Recommendations for Further Investigations

Scanning Facility Recommendations

With the pilot study coming to an end, it is NAL's intention that the National Agricultural Text Digitizing Project will become a full production program. For this program to be successful, some changes should be made to the central scanning operation; and in the future, bidding out work to commercial alternatives should be considered.

1. The page images, which can currently be displayed only on the high-resolution LaserView monitor, must be made available for display on any monitor. This change is critical if NATDP is to go into production. The discussions under workstation platforms and field experience make clear that the ASCII terminal displays for full text were unsatisfactory to users as tested, and that any workstation must be able to replicate the full bit-mapped image. There are several options:

    a. The images can be converted to another, more ubiquitous format. NATDP has done this on a small scale (two-hundred images) for two projects. It is time consuming, since the images must be converted one at a time to TIFF-5, and then converted to the desired format, usually in batch mode. It also requires a great deal of hard disk space, since each TIFF-5 image is about 1Mb. This is not a good approach when dealing with thousands of images.

    b. LaserData puts out software, called Add-Image, which purports to make their proprietary images available for display on any monitor. NATDP is attempting to locate a vendor for this software so it can be tested. Add-Image would have to be distributed with each CD-ROM to make it a viable option, but it has the advantage of requiring no other changes to the scanning workstation.

    c. Change the scanning workstation so that it outputs a more "standard" image format. NATDP would still require a format that is available on a high-resolution monitor, but one that could also be displayed and printed using any computer monitor. This option would require upgrading or replacing almost every piece of hardware in the scanning workstation, and then the system would have to be reintegrated. We would no longer have a workstation that could access the images on the pilot study disks, but this may be a necessary loss.

2. The text recognition software should be upgraded.

    The NATDP software is over four years old, dating to June 1988. Text recognition has improved significantly since then. For certain types of material, full text retrieval combined with selected graphics is still the best approach. Improving the text recognition software will allow NATDP to choose this approach when the material warrants it. The cost for software upgrade would be $7,000 to $9,000.

3. Permanent staff should be hired.

Hiring permanent staff would reduce the turnover rate, make the job more attractive, and give continuity between the various scanning projects. The equivalent of one full-time person would be adequate for scanning older materials that will not be converted to ASCII. For material that will require both scanning and text recognition, the equivalent of two full-time persons would be necessary.

4. A process should be developed for selecting material for scanning and for determining whether ASCII only, ASCII plus bit-mapped images, or bit-mapped images with fixed field retrieval elements will be employed. The discussion of error rates earlier in this report indicate that the type of material being scanned has wide variation in the quality of results. The cost of editing those items with high error rates precludes that alternative. Thus, certain kinds of material should only be scanned at the bit-mapped image level, with fixed field retrieval elements added to the record.

5. Specifications for archiving the data must be addressed, so that content is independent of the workstation platform and the specific distribution medium. The largest investment of a permanent production system will be the cost of creating the data; and that investment must be protected for the future.

6. NAL should continue to compare the capabilities of its in-house scanning facility to capabilities being developed by the commercial sector. Commercial options are becoming cost-competitive and could in the future become more cost-effective based on volume of scanning.

Field Test Conclusions

Overall acceptance of the various software packages tested for full text retrieval was dependent upon several factors: ease of use, quality of the graphics, and clearness of the screen displays. The overall response rate in the four test packages at no point reached more than one-third of the possible, making it difficult to attach statistical significance to the resulting data. Therefore, much of what can be reported are trends and personal observations of the participants in the tests. Many of the comments and responses to the questions indicate that a major problem is one often mentioned in the literature and seemingly inherent in the use of full text products: the inability of the user to see the relationship between the output and the original document. Some of the specific comments included: "needs some way to view an entire page on the screen", "would be nice if the menus, etc. emulated those of a major CD-ROM producer" (SilverPlatter, UMI, DIALOG, Wilson), "often didn't know where I was in the documents", "would be much easier just to flip through the pages of the originals", and "link needed to the bibliographic record". With an appropriate explanation and familiar frame of reference, users can be educated about unfamiliar, yet powerful features for searching full text databases.

Four key observations can be made from the field tests:

1. Libraries need to be very aware of the time required to participate in such experiments and to be sure that technical and professional support is provided.

2. Several sites never finished installing all the equipment or software because of lack of local technical expertise. For example, sites had to install CD-ROM driver software changes,

additional hardware or upgrades to hardware to support later software packages, and software under WINDOWS. The amount of technical support required from NAL was quite high. Even so, some sites could not successfully install all the hardware and software.

3. Hardware and software were evolving rapidly during the period of the research and demonstration project. As software packages improved, they often required more powerful hardware than participating sites had been told they would need. It was difficult for many test sites to fund upgrades to equipment to utilize the newer software packages. This need for constant upgrading of hardware and software will confront all libraries in the future as they are faced with providing access to increasing numbers of electronic publications and databases.

4. Librarians initially expected the full text disks to operate like bibliographic CD-ROM services and were frustrated that they did not. The project educated participants about the complexities of full text formatting and searching. It is another level higher of complexity from citation databases, both in design for clarity of use and in the ability of the user to search and manipulate text.

Other Recommendations

1. By the end of the project, it was clear that dedicated workstations were less than ideal and that a client-server model whereby users could gain access from personal computers through networks was desirable. In addition, users wanted the ability to download from the disk to their workstation and/or to a server from which they could download and print. This architecture is a critical piece of Phase 3 of the project, under the leadership of North Carolina State University.

2. Vendors and librarians need to pursue licensing solutions that support the need for networked access to CD-ROM collections. Current licenses often restrict access to a specific workstation or increase fees for networked applications in ways not directly related to actual use or patron needs.

3. The largest investment facing publishers and libraries in making full text image databases available is the cost of creating and/or converting the data. They must be able to ensure the integrity of the data or content regardless of hardware and software changes. Thus it is imperative that standards for bit-mapped images be established and accepted in the marketplace as soon as possible. Lack of accepted standards for archiving the data will inhibit the growth of this technology, particularly for purposes of preservation.

4. In addition to establishing standards for archiving the data, the field must find ways to ensure that master disks or the source databases are protected and refreshed as the electronic data ages. Producers, publishers, and libraries have no established responsibilities or expectations for protecting and archiving master disks.

5. NAL must utilize software that supports multiple platforms (DOS, OS/2, Macintosh, UNIX operating systems), as the user base of hardware must be relied upon for distribution. Proprietary hardware and software will inhibit acceptance and use of electronic publications. While helpful in prototype, the use of proprietary hardware and software is a dead end in production.

Appendix 1

Participating Libraries

Auburn University, Auburn, Alabama
University of Alaska, Fairbanks
University of Arizona, Tucson
University of Arkansas, Fayetteville
University of California, Berkeley
University of California, Davis
Colorado State University, Fort Collins
Delaware State College, Dover
University of Florida, Gainesville
University of Georgia, Athens
Fort Valley State College, Fort Valley, Georgia
University of Hawaii, Honolulu
University of Idaho, Moscow
University of Illinois, Urbana-Champaign
Iowa State University, Ames
Kansas State University, Manhattan
University of Kentucky, Lexington
Louisiana State University, Baton Rouge
University of Maine, Orono
University of Maryland, College Park
University of Maryland - Eastern Shore, Princess Anne
University of Massachusetts, Amherst
Cargill Information Center, Minneapolis
University of Minnesota, St. Paul
University of Missouri, Columbia
Montana State University, Bozeman
University of Nevada, Reno
University of New Hampshire, Durham
Rutgers University, New Brunswick, New Jersey
New Mexico State University, Las Cruces
North Carolina State University, Raleigh
North Dakota State University, Fargo
Ohio State University, Columbus
Ohio State University, Ohio Agricultural Research and Development Center, Wooster
Oklahoma State University, Stillwater
Oregon State University, Corvallis
Pennsylvania State University, University Park
University of Puerto Rico, Mayagüez
Clemson University, South Carolina
University of Tennessee, Knoxville
Texas A&M University, College Station
Prairie View A&M University, Prairie View, Texas
University of Vermont, Burlington
Virginia Polytechnic Institute, Blacksburg
Washington State University, Pullman
University of Wisconsin, Madison

Appendix 2

Advisory Panel

Noreen Alldredge, Montana State University
John Beecher, North Dakota State University
Paul Gherman, Virginia Polytechnic Institute
H. Joanne Harrar, University of Maryland
Charlene Mason, University of Minnesota
Thomas Shaughnessy, University of Missouri
Clifford Lynch, University of California System

Appendix 3

Library Project Staff

National Agricultural Library

Pamela Q. J. André
Judith Zidar
Dan Starr
Gary McCone
Carol J. Shore
John Stetka

University of Vermont

Nancy L. Eaton
Jerry V. Caswell
Albert Joy
Mara Saule
Linda McDonald
Reference Department staff

Iowa State University

Nancy L. Eaton
Eleanor R. Mathews
Diana Shonrock
Gerard McKiernan
Reference Department staff

Appendix 4

Strategic Considerations in Platform Selection for the Production Phase of the NAL Text Digitizing Program

Original Report

Strategic Considerations in Platform Selection for the Production Phase of the NAL Text Digitizing Program

Clifford A. Lynch

December 31, 1989

Introduction

This white paper outlines for the NATDP Steering Group some of the strategic issues that I believe must be considered in the selection of target workstation platforms, as the Text Digitizing Program proceeds from an experiment into an ongoing production operation. I believe that the discussion results from these points should be factored into the selection process for a "production" CD-ROM retrieval package as we move towards ongoing operational status.

Basic Philosophy

During the prototype test and evaluation phase of the NATDP, it was necessary to procure special-purpose workstations that could display and print high-resolution bit-mapped images of material being published by the project on CD-ROM; such workstations were not part of any "standard" market offering and no standards existed in this area.

During the three years of the prototype test and evaluation phase of the project, the marketplace environment has changed significantly. Standards now exist in some key areas, and more are being developed. Standard commercial offerings have expanded significantly in the workstation marketplace. One very real difficulty that still faces the NATDP in planning future actions is that the technology available in the CD-ROM software package marketplace has not kept pace with changes in the hardware and operating system base. At least in the short term, this constrains choices; however, NAL and the Text Digitizing Project have an opportunity to exercise further leadership in this marketplace and to provide encouragement for key vendors to accommodate the existing and developing standards.

I believe that the guiding principle for the operational phase of the Text Digitizing Project should be to use the existing, in-place hardware base wherever possible in participant libraries. This existing hardware base consists of the following elements:

- PC's running DOS (these are the trailing edge)
- PS/2 machines running OS/2
- PS/2 machines running AIX (IBM UNIX)
- Macintosh machines running the Apple operating system
- Macintosh II machines running AUX (Apple UNIX)
- Various types of workstations running UNIX variants

I feel it is essential that we move away from a non-industry-standard SAIC platform for the production phase of NATDP, for several reasons. The SAIC systems are not compatible with other CD-ROM products on the market. SAIC is at best expensive, and at worst impractical, as a nationwide hardware and software supplier. The lack of competition for the specialized SAIC hardware/software base will keep prices high and will slow development.

In this discussion I explicitly exclude networking issues. The entire question of network servers and network transmission of document images is still the subject of an active research and development project within the NATDP. It may be at least another year before future directions for network service to workstation populations have been validated through the experimental efforts being carried out under NATDP auspices.

Major Issues

The SAIC bit-mapped workstation was unique in two areas: the raster display itself, and the high-bandwidth laser printer interface, which permitted relatively rapid printing of bit-mapped images. Both of these were areas where the base of industry-standard workstation technology could not meet NAL's needs during the prototype phase of the project. In addition, use of the SAIC workstations provided a controlled environment that was valuble during the evaluation phase.

Currently, there are both vendor-specific and vendor-independent standards for handling high-resolution displays. Microsoft windows for the IBM PC-based platforms and the native Apple Macintosh support for the Machintosh are vendor-specific. In the vendor-independent area, there is the X Window system, which came from MIT's Project Athena, funded by IBM and Digital Equipment. The X system is implemented not only on most UNIX platforms but also on MS-DOS and the Macintosh operating system. An OS/2 version may be available by early 1990. Ideally, any future software should support X; as a backup position, it should support at least X as an option, Microsoft Windows, and the Macintosh system. A final consideration is OS/2 Presentation Manager in the PS/2 area. However, it seems likely that some convergence will occur between Microsoft Windows and Presentation Manager in 1990. This area should be carefully monitored.

The second problem area during the prototype phase was the interface to laser printers. Again, de facto standards have emerged during the years covered by the prototype period — in this case Adobe PostScript. Virtually all laser printers now support PostScript. Here, however, more than in display management, there is a legitimate performance issue. Normally, PostScript is used for typesetting — font information and layout data, rather than pure bit maps, are passed to the printer. Transmitting an entire page image as a bit map is still quite slow, due to limited workstation/PC bandwidth. However, trading off the ability to operate on a very wide range of platforms for print efficiency seems somewhat shortsighted. I feel that NATDP would do better accepting a certain degree of performance penalty in this area rather than compromising the ability of its disks to work on a wide range of platforms. If necessary, a high-performance nonstandard option could be added for print-intensive situations — but this should be an exception, not the standard operating package.

Summary: Proposed Selection Criteria

The ISO 9660 CD-ROM labeling standard supports boot blocks on a CD-ROM for multiple machine types. I would propose that we consider the following selection criteria for CD-ROM retrieval software:

1. *Required*: support for OS/2 and Macintosh operating system

   Desired: support for PC-DOS, UNIX (IBM AIX, IBM AUX, etc.)

   The vendor should allow mastering of a CD-ROM that contains software for multiple platforms at no extra cost.

2. *Required*: support for bit-mapped display under Macintosh and OS-2 (Microsoft windows)

   Desired: support for OS/2 Presentation Manager and X windows

   — or —

   support for X windows under Macintosh, OS/2, PC-DOS, AIX, AUX and other UNIX variants (probably the preferred long-term approach)

3. *Required*: support for Postscript Printers

   Desired: high-performance special printer interface option

Appendix 5

NATDP Identification of the Retrieval System of the Future

Final Report,

Sections I and II

NATIONAL AGRICULTURAL TEXT DIGITIZING PROJECT (NATDP)
IDENTIFICATION OF THE RETRIEVAL SYSTEM OF THE FUTURE

FINAL REPORT

PAULINE A. ZOELLICK
BOULDER, COLORADO
11 OCTOBER 1990

THIS REPORT CONTAINS CONFIDENTIAL INFORMATION

SECTION I: THE WORK

STATEMENT OF WORK AND WORK PRODUCTS

The purpose of this study was to identify a software retrieval system for use by the National Agricultural Text Digitizing Project (NATDP) in the development of future text and image products on CD-ROM.

Since 1987 the NATDP, in a cooperative project involving the National Agricultural Library (NAL) and forty-five land grant libraries, has been involved in the capture of full-test and images in digital format for publication on CD-ROM.

As a part of Phase I of that project, three different retrieval systems were used in the preparation of document collections.

Based on these activities of the NATDP, it was decided to conduct an in-depth assessment of currently available software packages before selecting one for use in development of products in the next stage of the project.

A Statement of Work dated May 16, 1990 was prepared containing a checklist outlining NATDP's requirements for a software system, a questionnaire listing optional desired features, and a list of retrieval systems and their respective vendors.

A consultant was to review the retrieval and build software packages and select three to recommend to NATDP. Final selection of a retrieval system was to be made by an NATDP Software Evaluation Team.

The evaluations were to be based on actual products and product prototypes on CD-ROM, not demonstration floppy disks or magnetic files.

The consultant was also to collect information as to financial status and business practices of the final three vendors recommended to NATDP.

Work products were to include a checklist of the assessment for each package reviewed, and for the top three, a written report on each , a completed checklist, and a completed questionnaire of optional features.

The NATDP Software Evaluation Team was then to meet at NAL in Beltsville in mid-October 1990 to select the final package for use by the NATDP in the next stage of product development.

METHODOLOGY

Supplied by NATDP

NATDP supplied a list of twelve retrieval system vendors reported to have products capable of meeting NATDP's requirements. (Appendix A: 7/2/90 Contacts for Full-Text Retrieval Software Vendors)

NATDP supplied a checklist of system requirements and a questionnaire of optional features. (Appendix B: National Agricultural Text Digitizing Project, CD-ROM Retrieval Software Evaluation Tool and Appendix C: Sample Questionnaire)

NATDP supplied an Hitachi Model CDR-1503S CD-ROM player and board and the Microsoft MS-DOS CD ROM Extensions Version 2.02 with device drivers as distributed by Meridian Data, Inc.

Supplied by Consultant

The checklist and questionnaire were expanded and combined into a database containing:

- required features checklist items, original checklist numbers, indicated "*"
- desired features questionnaire items, in Section Q--Questionnaire
- features added by the consultant, features numbered #X-Z
- feature descriptions for all items

Over the course of the evaluation process data were added to this database on each software system reviewed. These data were then used in evaluation and report production.

The database was used to produce various work forms including:

- a list of feature names and descriptions for use by the consultant, the vendors and NATDP Software Evaluation Team to assure that the same meanings were attached to the feature names by all involved in the evaluation. (Appendix D: NATDP Software Review--Feature Names and Descriptions)
- blank evaluation forms for use by the consultant and by the NATDP Software Evaluation Team in recording system features and capabilities (Appendix E: NATDP Software Review--Blank Evaluation Form)

A list of abbreviations used in recording retrieval system features was prepared. (Appendix F: NATDP Software Review--Key to Abbreviations Used)

Also prepared was a list of factors to be considered in evaluating vendors and software systems beyond the system features listed by NATDP. (See Section II: Elements of Software and Vendor Evaluation, Factors for Consideration in Selecting a Vendor)

The consultant's computer system was used to evaluate products. It consisted of an IBM-compatible manufactured by DTK with a 10 Mhz clock, 640K RAM extended to 2MB, a

40MB hard drive, one 3.5 and one 5.25 inch floppy drives, an EGA monitor, and a Hewlett Packard LaserJet III laser printer. To this was connected the CD-ROM player provided by NATDP.

Procedures Followed

The original set of twelve vendors and retrieval systems (in alphabetical order by vendor):

Vendor	Retrieval System
Dataware Technologies	CD Answer
Executive Technologies	Search Express
Fulcrum Technologies	Ful/Text
Knowledge Access International	KAware 2
Knowledge Set Corporation	KRS
Nimbus Information Group	ROM Ware
Personal Library Software	Windows Personal Librarian
Quantum Access, Inc.	Quantum Leap
Reference Technology, Inc.	Reference Book
Retrieval Technologies	Re:Search
TextWare Corporation	TextWare
TMS	InnerView for Windows

All twelve vendors were contacted by the consultant in order to do an initial screening on the basis of NATDP requirements.

As a result of this first round of screening, eight vendors were selected for further review:

Vendor	Retrieval System
Dataware Technologies	CD Answer
Executive Technologies	Search Express
Knowledge Access International	KAware 2
Knowledge Set Corporation	KRS
Personal Library Software	Windows Personal Librarian
Quantum Access, Inc.	Quantum Leap
Reference Technology, Inc.	Reference Book
Retrieval Technologies	Re:Search
TMS	InnerView for Windows

Fulcrum Technologies' Ful/Text was eliminated at this stage. Both Fulcrum and the consultant considered that NATDP's needs would be better met by dealing with a Fulcrum licensee, such as Reference Technology, to acquire Fulcrum application level software. Contracting directly with Fulcrum was recommended only if NATDP were to be interested in use of the Fulcrum search engine in building NATDP's own application level software.

Nimbus' ROM Ware and TextWare's TextWare were also eliminated at this stage. They were found to be oriented toward highly structured data and as such were not able to support the full-text and image requirements of NATDP. (See Section IV: Report on Non-Finalist Systems)

The nine vendors selected at the end of round one were contacted by Michael Cramer of Virginia Polytechnic Institute & State University Library to solicit product for the consultant's review.

The products sent by the vendors to the consultant were then installed on the consultant's computer and reviewed. Calls were made to the vendors if/when appropriate for installation instructions, clarifications, questions, etc.

Data were obtained by the consultant on all systems through vendor interviews, review of vendor documentation, review of marketing materials and actual product use and evaluation.

These data were summarized and entered into the database described above.

The three products selected for final review by the NATDP Software Evaluation Team were (in alphabetical order by vendor):

Vendor	Retrieval System
Knowledge Access International	KAware 2
Personal Library Software	Windows Personal Librarian
Reference Technology, Inc.	Reference Book

Each finalist vendor was sent the list of NATDP Vendor Selection Feature Names and Descriptions and a listing of the data for their particular system as generated by the consultant. The vendor verified, corrected or added data and returned the list to the consultant. Corrections were incorporated into the database.

A matrix was prepared to show all three of the finalist systems against the evaluation criteria. Written reports were also prepared describing each of the three. (See Appendix K: Evaluation Matrix--Final Three and Section III. Reports on Finalist Systems)

A matrix was prepared showing the eight non-finalist vendors evaluated and the key elements of the evaluations up to and including the point at which each vendor system was eliminated from the final round. (Fulcrum having been eliminated in the first round, this eight plus the finalist three represent all systems reviewed.) (See: Appendix L: Evaluation Matrix--Non-Finalists and Section IV. Report on Non-Finalist Systems)

It was decided in consultation with NATDP that information on all non-finalist systems would be more valuable to the project than completed checklists for only some of the systems as originally planned. Data is thus provided for all systems reviewed. The information supplied for the non-finalists varies from vendor to vendor dependent on the point at which the package was determined to be unsuitable for NATDP's needs.

A table was constructed to show both build and user license costs across the three finalists. (See Section III. Licensing Information)

The final three vendors were also asked to supply customer contact information and financial information and/or credit references.

The three finalist vendors were contacted by NATDP and invited to present both business and technical information on their systems at a week long meeting of the NATDP Software Evaluation Team in Beltsville, Maryland October 15-19. Each of the vendors was given a full day in which to present information, do demonstrations and answer evaluation team questions.

At the beginning of the week of evaluation Members of the NATDP Software Evaluation Team were to be supplied with:

- NATDP Software Review--Key to Abbreviations Used (Appendix F)
- NATDP Vendor Selection--Feature Names and Descriptions (Appendix D)
- NATDP Vendor Selection--Blank Evaluation Form (Appendix E)
- Elements of Software and Vendor Evaluation (Section II)

To be available to the Team members if they so desired were:

- matrix forms for each of the top three vendors
- matrix form for all of the top three vendors (Appendix K)
- textual summaries of each of the top three systems (Section III)

Additional documents prepared for the NATDP Software Evaluation Team were:

- Licensing Information (Section III)
- Client Reference Checks (Section III)
- Credit Reference Checks (Section III)
- textual summaries of the non-finalist (Section IV)

Evaluation Methodology Employed

Software packages were evaluated, using as a basis, NATDP's requirements as set forth in the documents supplied by NATDP. In addition items were added to the evaluation criteria by the consultant as packages were reviewed and as interviews with vendors proceeded. The items chosen for addition focused on newer features of systems, additional clarifications of some features, etc. Items were added only if they represented necessary clarifications or features that were felt to be essential to an informed decision on the part of the NATDP Software Evaluation Team.

Retrieval systems were installed and used, first without reference to printed documentation, and then with the assistance of manuals, "guided tours", search tips, etc. as supplied by the vendors. Build software, if supplied, was evaluated primarily from the documentation.

Clarifications, installation assistance, and other routine questions were asked of the vendors if this was necessary or seemed reasonable. For example, where products produced for specific clients were supplied for review, clarification was sought as to which features were part of the software and which were client imposed designs.

Telephone interviews were conducted as time allowed with client contacts provided by vendors. Clients were selected for calling based on the similarity of their application(s) to those of NATDP, the length of their association with the vendor, the developmental nature of their association with the vendor. Client interviews were summarized and are presented below. (See Section III. Client Reference Checks)

Credit references were called or faxed and information received from them is summarized below. (See Section III. Credit Reference Checks)

Information known to the consultant in regard to business practices of the respective companies was also taken into consideration.

In collecting, analyzing and providing information to the NATDP Software Evaluation Team the primary concern was in identifying well structured, reliable, and forward-looking build and retrieval software, produced and supported by responsible, financially viable companies.

The information presented in Section II outlines key elements of that selection process as it expanded upon the requirements originally set by NATDP.

SECTION II: ELEMENTS OF SOFTWARE AND VENDOR EVALUATION

FACTORS FOR CONSIDERATION IN SELECTING A VENDOR

General

There is much to be considered in selecting a CD-ROM software system that cannot be expressed simply be listing software features, or by looking at user interfaces.

You are selecting an information seeking and handling concept that has been developed and implemented in large part without your input. The perceived value of your information will be influenced greatly by the face it presents to the world--a face shaped in large part by the software vendor.

There is no single best way to represent information. Different types of data need different processing and presentation methods. Different users have different information needs and information seeking habits. Systems are usually built with a specific type of data and/or user in mind. Many systems now available also allow a great deal of flexibility in configuring a system to meet your specific needs.

You are contracting *with* a company as well as *for* software. Everything that your product does will be influenced by the vendor's development schedule. So, before choosing a system, it is wise to understand as much as you can of the system and its creators.

Robustness of Code

Software code that has not been adequately engineered, implemented and/or tested can have a tendency to "blow up", do unpredictable things and/or freeze the system. This is not acceptable behavior unless this is code still in the early testing stages. If so, the vendor should make this clear to you before you begin to work with it.

Ability to Configure Software Applications

Look for a package that expects you to want to configure the interface to suit your data and your customers. Build software, for example, should allow you to change labels, screen layout, order and content of displays.

These are now considered standard options, and you need this type and degree of control over software that you intend to use for multiple applications.

Platforms-Minimum and Other

It is difficult to configure code that works well under a broad range of hardware configurations. NATDP is looking for software that will produce and display high resolution images and yet be usable on an installed base of low resolution monitors as well.

High resolution images can usually be displayed on low resolution monitors, but with loss of readability due to the limited display resolution and therefore space on the screen. It is generally possible to have the software zoom the images to an acceptable level for examination, but this then requires that the user pan around the page to seek information. While this may be a bit inconvenient, it is for most an acceptable solution, as it gives high quality images to those with equipment and utility to those without it.

Look for systems that can already handle both.

Customer Base

Look for software that has been used to produce products similar to those you want to produce.

Most systems are optimized around a certain type of data (full text, directory data, bibliographic records, images, etc.) and although they can be made to work with a variety of data they don't necessarily do it well.

Integral Software

Some products require that you also purchase another software system to use or manage them. For example, Windows and GEM provide environments for the software.

Where there is or will be a large installed base of the software it can work for you in providing common conventions and a look that is familiar to a customer. If the customer already has the environment software, your product then has the effect of leveraging the customer investment. If, however, the software is not already in place in your market it can, in effect, raise the price of your product. Feedback from your potential market is important here.

General Navigational Features

By virtue of the sheer amount of data available and the newness of the media, one of the most common problems with any CD-ROM based product is difficulty in keeping track of what is on the disc and where a user is at any given time. Both of these problems can be helped by software that suits the data and can be configured to provide "navigational" aids to the user.

The display of data from within the file, the ability to view data and image simultaneously, the ability to step through the data in a sequence that makes sense to the user, the ability to use screen labels that reflect the terminology used in the users vocabulary, all help in this regard.

Image and Text Display

Systems that allow the simultaneous display of text and images assist the user in maintaining control of the work environment. Look for systems that provide for simultaneous display.

At minimum, directions for using the images should be easily available on the screen while the image is being displayed.

Preservation of Input File

Retrieval system houses are beginning to develop software that can use already tagged data with display and retrieval "filters", rather than requiring that you put your data into their "standard" form.

This allows you to retain structural information that is of value to you on the CD-ROM while also taking advantage of the vendor's retrieval package. Use of your structural tags also allows you more display and navigation options that suit your data. Look for systems that are using structural information in this way or are planning to do so.

Contract Services

If you are at all likely to want or need assistance in designing your product, preparing or processing your data, or training your staff, look for vendors who are set up to provide a range of such services. Some do this as a matter of course. It provides options as you move into new areas, have staff shortages, are preparing difficult data, etc.

Some firms concentrate primarily on software development but will provide such services on request. They are rarely structured and staffed to do so on an ongoing basis, however.

Features and Functionality

Beware the feature count. If these are not features you need, or if they do not provide real functionality, you may be sacrificing robustness or quality in other areas.

Be sure to evaluate trade-offs between "ease of use" and search and display functions that are needed by your user group. At the same time beware overpowering the system. Again, configurability counts.

Non-Issues: Large Image and Text Files, In-House Build

During the early days of the NATDP finding software that could handle large files of text and images and had documented in-house build software was a significant problem. Now, there are many firms who can provide such products. But still few who can offer all of what NATDP has asked for. Some of the search and display features requested by NATDP are considered advanced. Be sure your users need them.

License Costs: One Time vs. Repeat vs. Functionality

Prices and licensing structures vary between vendors, so it is well worth considering what this means in terms of functionality, robustness, staying power and support.

Vendors offer quantity pricing and often will also make allowances for the visibility of the client and the opportunity this represents for referral business.

Speed

Most retrieval packages evaluated during this review had fast and effective retrieval engines. Some firms license only their retrieval engine and let you build your own applications around it.

Although this may not be something that you are interested in now, the option can be indicative of the quality of the engine and of the general level of expertise and technical support available within the firm.

Resources: Design, Technical, Support

Look for firms that can provide experienced staff in sufficient numbers to allow you to complete your products.

A staff with a broad range of capabilities will provide the internal checks that generate marketable, supportable packages.

Appendix 6

Roles for Telecommunications and Computer Networking in the National Agricultural Library Text Digitizing Project

Roles for Telecommunications and Computer Networking in the National Agriculture Library Text Digitizing Project

Clifford A. Lynch
December 1, 1988

Executive Summary

This report discusses roles for telecommunications and computer networking in furthering the objectives of the National Agriculture Library's Text Digitizing Project. Both architectural and technical system design issues are considered. The report also examines the role of networking in interlibrary loan and demand document delivery for the NAL collections, and the possible uses of the network to support other database access and intersystem linkage initiatives being pursued by NAL and the Department of Agriculture as a whole.

The report's primary recommendations are:

• NAL is best served by participation in the evolving national research internet rather than by the construction of its own private network (although, if sufficient funds are available, a case can be made for the development of a specialized NAL network for document distribution on technical grounds). There are a number of related policy decisions identified in the report which should be made prior to any further actions involving networking.

• NAL should take an active role in the policy and standard issues involved in the development of the national research internet.

• NAL should institute a project to install a connection to the national research internet, and exploit the internet in subsequent projects.

• NAL needs to carefully evaluate whether use of electronic document delivery for interlibrary loan/demand document delivery is really cost justified in terms of service improvement to its user community. Request processing at NAL rather than the time required to physically deliver the document in question may be the limiting factor in service improvement.

• NAL should give serious consideration to standardizing on X windows to provide hardware independence for bit mapped display platforms at the conclusion of the current round of prototypes being developed as part of the Text Digitizing Project.

• The next phase of the Text Digitizing Project should investigate distribution of document text and images over local area networks and regional internetworks.

1.0 Introduction

The National Agriculture Library (NAL) has embarked on a major project to capture collections of agricultural documents in digital form (either as page images or as ASCII text) and to make these collections available to a broad user community through the use of advanced computer-based technologies. The first phases of this project have dealt with the problems of data capture and the publication of collections on optical media (CD-ROM).

As the project grows and matures, it is clear that the technologies of telecommunications and computer networking can play a number of possible roles in making the digital document collections accessible to NAL's various user communities. These technologies can serve as a complement to the current optical publishing approach, as a parallel means of providing access to the information, or as a complete substitute for today's optical publishing activities. In addition, networking will play an important future role in the way the data is used by the members of NAL's user community; consequently, an understanding of this aspect of networking will provide important guidance in the evolution of the optical publishing effort.

The aim of the Text Digitizing Project is to distribute material that is expected to be of relatively wide interest to the user community. For the foreseeable future, only a small part of the holdings of the National Agriculture Library can possibly be published through the Text Digitizing Project. This is due in part to budgetary and time constraints. Even beyond these limitations, however, it does not make sense to publish the library's entire holdings. Much of the information held by NAL is called for only infrequently. When such information is needed, NAL serves as a national resource to satisfy interlibrary loan requests.

Telecommunications and networking technologies offer potential service improvements and cost reductions for interlibrary loan requests that fall outside the scope of the material being digitized through the Text Digitizing Project and, in this sense, can serve as a complement to the publishing approach being taken by the project. For some types of material, in fact, the publishing and interlibrary document-transfer functions can work together: It is possible to publish very large numbers of bibliographic citations and/or abstracts (in ASCII form or even, for the abstracts, as images) on optical media; these abstracts can be used as the basis for initiating interlibrary loan requests resulting in the transmission of document images digitized on demand. Existing databases such as AGRICOLA, available both online from service bureaus and as CD-ROM publications, already provide substantial coverage of the literature of agriculture.

NAL's role as a national resource within the agriculture community has recently received new emphasis, focus, and recognition through initiatives such as Director those described in Joseph Howard's paper entitled "National Agricultural Library and Information Network: Outline of Plan." While the concept of a network, as presented in that paper, is of a logical network of participant organizations cooperating in the sharing and dissemination of information, Director Howard's plan does call specifically for infrastructure development as an essential step toward the realization of the logical network. Telecommunications and computer networking are essential parts of this proposed infrastructure.

Such an infrastructure, once created, can serve multiple roles in support of U.S. agriculture. The Cooperative Extension system is working with NAL to exploit the potential of national databases, electronic technologies, and expert systems. Work is underway to link NAL and selected Agricultural Research Service libraries; the existence of a network would facilitate this project. A longer-term NAL goal is to make its new online public access catalog available for remote access; again, such an undertaking could be helped by the presence of a network linking remote users to NAL. Appropriate information distribution and shared communications channels implemented through telecommunications and networking technologies can potentially play a key role in such initiatives as they develop.

1.1 Overview

This report has several objectives. First and most important, considers the various roles that telecommunications and computer networking can play in information distribution and access. This is done in section 2, without a great deal of emphasis on the details of particular telecommunications and networking technologies. In the end, choice of appropriate technology will be dictated by the architecture and function of the information delivery and access systems. Next, section 3 of the report examines a range of available technologies, discussing their status, their economics, and how they might fit into the various possible designs for delivery and access systems.

Section 4 of the report considers the ways in which telecommunications and networking issues relate to how the institutions that make up NAL's user community provide access to the data. This is a complex and increasingly important topic. The first phases of the Text Digitizing Project conceived of the optical publishing effort as placing copies of databases at institutions within the user community. These institutions would then use stand-alone workstations to access the databases they had received. In fact, however, many of the institutions within the user community have or are developing complex internal networks to serve people within their organization. At many institutions, the people who need access to the databases are geographically scattered at multiple sites. (In extreme cases, such as Hawaii, users are on several islands.) Even when the participant institution is located at a single site, such as a university campus, end users of the databases want the convenience of accessing these databases from their offices, labs, and homes. Finally, many of the participant institutions are actually the lead organizations in local consortia and serve one or more secondary institutions. Local area networks (LANs) at these institutions have become interlinked in complex ways as interinstitutional internetworks have developed.

Thus, at least potentially, many of the members of the user community would become redistributors of the databases, both for their own internal purposes and to other institutions within the overall NAL user community. This environment leads to new possibilities for NAL to distribute data by using telecommunications and computer networking. One can envision hierarchical, regional centers that obtain data directly from NAL and then redistribute the databases or provide database access for the remainder of the user community.

As the project matures, I believe that the key networking issue will become the development of retrieval systems that can operate in a local and regional network environment.

Section 5 considers applications of networking and telecommunications technology to demand document delivery as part of NAL's roles as a national resource within the interlibrary loan system in the United States. This section focuses on material that will not be published as part of the Text Digitizing Project. Particular emphasis is given to investments in technology and infrastructure that can serve multiple purposes by supporting both the Text Digitizing Project and NAL's general interlibrary loan activities.

The final section of the report makes some recommendations for the use of telecommunications and computer networking technologies in NAL's Text Digitizing Project and related activities. In addition, it raises some management, financial, and organizational issues that need to be considered along with these technologies, particularly as they apply within the broad context of a National Agricultural Library and Information Network as proposed by Director Howard.

1.2 Terminology: Telecommunications and Networking

This report differentiates between a telecommunications link and a network. The distinction is very important. A telecommunications link (a phone line, a microwave link, a satellite link, and so on) is simply a means of moving bits--zeros and ones--from one end of the link to the other. Satellite links are peculiar in that they have many endpoints; using a satellite link, it is possible to ship bits from one site to many sites in parallel.

A network (such as a packet-switched computer network) is a much more complex system that provides functionality oriented toward the end user and incorporates one or more telecommunications links together with some intelligence. For example, a network may worry about such issues as routing of data, error correction, recovery from telecommunications link failures, and delivery of information to specific addresses.

A telecommunications network is a collection of links interconnected by switching and routing logic; it may carry voice, digital data, video, or a combination of these signals. A computer network is a set of computers that are interconnected by a telecommunications network; a computer network is designed specifically to support the exchange of digital data among the computers. It has become increasingly common to find that the telecommunications network supporting a computer network is designed to carry digital data, rates then employing devices such as modems to transport digital data over analog trunks.

Note that a computer network is designed to link computers, not to provide terminal access to remote computers. In a true computer network, terminals are connected to a computer rather than directly to the network, and access to network resources is obtained through the local computer to which the terminals are connected. The local computer may be a general-purpose processor with significant local computing power, or it may be a small, special-purpose machine (a terminal server) designed only to provide its terminal population with network access.

To distribute data to a number of sites, a network of some kind is needed. This requires telecommunications links, but it will requires some intelligence, housed either in the computers attached to the network or in dedicated switching computers embedded directly in the network, separate from the machines running the actual applications.

Confusingly, within the library community, the term "network" has traditionally been used to describe an organizational structure consisting of a consortium or other grouping of libraries and related information resources. One example is the set of regional networks that have traditionally brokered OCLC services. Such organizational networks may or may not actually be supported by a telecommunications or computer network that allows communications and information interchange among their members. In this report, the term "network" will normally refer to an actual, physical network of computers linked by telecommunications facilities.

1.3 The NAL Context

Before the potential roles of telecommunications and networking technologies can be considered, two essential background elements must be described: the nature of the user community and the nature of the data involved. These will play a central role in determining the desired functionality of distribution and access systems and in defining the viability and cost of candidate technologies.

1.3.1 The NAL User Community

Very broadly speaking, the user community can be divided into three segments:

- The direct, immediate user community. This is made up of the land-grant libraries and other institutions currently participating in the Text Digitizing Project.

A considerable amount is known about the situation at these institutions. A survey of this group conducted in late 1987 by the Text Digitizing Project included the collection of information about existing and planned network facilities at each site. Although there is some reason to believe that the results of this survey underestimate the amount of network connectivity available at the participant sites (perhaps because the people responding to the survey were not fully informed about network links available at their institutions, possibly never having had cause to use these links), the 27 returned questionnaires indicate the following:

Twelve sites have LANs; six more have them in the planning stages. Based on my knowledge of the institutions in question, I believe that at least two of the sites that claim they don't have LANs probably really do, but these networks may not go to the library. Thus, we can expect that at least 75 percent of the participants will have a local area network in place within about a year.

Three sites indicate that they are connected to the national internet. I believe that an additional two or three sites are actually connected, although the respondent was unaware of this connection. (At many universities, the first department that gets a national connection is the electrical engineering or computer science department. Often this connection remains a departmental resource for the first few years of its existence, until the rest of the campus becomes aware of it. Also, in some cases the funding agencies providing the link to the national network place restrictions on the use of this link.) Thus, we can assume that within a year at least a quarter of the participant sites will be connected to the national internet. Given the rapid increase in connectivity that is taking place across the country, I would speculate that this group will actually approach 50 percent of the participants by, say, the middle of 1989.

- The broader U.S. user community. This includes more than 100 land-grant libraries and other institutions scattered throughout the continental United States, Puerto Rico, Guam, Hawaii, Micronesia, Alaska, and other locations. The institutions currently participating in the Text Digitizing Project are a subset of these.

Many other organizations within the United States would also benefit from the data being captured and distributed by NAL. Such groups include other universities, research institutions, government agencies (federal and state), and the agribusiness community. At least potentially, these institutions also form part of the broader U.S. user community.

- The international user community. Internationally, a wide range of governmental, educational, and commercial entities may be interested in the data being captured and distributed by NAL. International telecommunications is highly regulated, and international computer networking is extremely sensitive and political. Thus, international networking and telecommunications issues involve substantial additional complexities and uncertainties.

This report focuses on the first two segments of the user community. Issues unique to international telecommunications and networking are mentioned only in passing. To some extent, the Western European nations, Australia, New Zealand, and Japan participate in the U.S. networking infrastructure; by and large, the remainder of the world does not.

1.3.2 The Document Collections

The nature of the document collections being captured and distributed by NAL as part of the Text Digitizing Project will play a key role in determining what types of telecommunications-based access and distribution systems make sense. Some key questions are as follows:

- How much does the collection change over time? Is new material added? Is old material corrected?

- How important is it that the user have access to the latest information (for example, updates and additions) in a nonstatic document collection? Is six-month-old information adequate? Or must updates be provided monthly, weekly, or even daily? A properly designed telecommunications system could potentially deliver new or updated information to the user within minutes of its capture. Such a system is likely to be quite expensive, however. Is the cost justified?

- How quickly must queries against the document collection be satisfied? Does the user need documents while he or she waits (essentially, within a few minutes of submitting a search), or is overnight (or slower) document delivery sufficient?

- Will the documents be delivered exclusively as ASCII text, primarily as ASCII text with a limited amount of image information, or as digital images of printed pages? This is a key question that we hope the optical publishing pilot experiments will resolve. Documents are initially being captured as images and are then converted to ASCII text by computer algorithms; we have yet to determine whether the error rate in this conversion is acceptable and to what extent the loss of detailed typographic and layout information degrades the usability of the captured documents. Finally, we have yet to learn what percentage of the document collections must be maintained in image format due to the presence of illustrations, diagrams, equations, and similar material.

Delivery format is a particularly crucial issue. If the documents are almost always delivered as ASCII text, then a relatively slow-speed telecommunications approach may be feasible. However, if images are the predominant delivery format, a broad-band, and thus potentially very costly, telecommunications approach will be required.

As a further complication, it should be noted that the various document collections that NAL will be capturing are likely to be quite heterogeneous in their characteristics. Some may be relatively static whereas others will be volatile, with a significant value to current information; some may be almost entirely text material whereas others may have to be handled primarily as image-format collections due to problems in omnifont recognition or the presence of extensive illustrations, tables or equations.

The considerations for documents transmitted in response to interlibrary loan requests rather than ones that form part of a collection that is formally distributed or published through the Text Digitizing Project are quite different:

- For documents transmitted on demand, images rather than ASCII text are the natural format for distribution. The need for an editing cycle for ASCII text would greatly reduce responsiveness to such interlibrary loan requests. Because the volume of information will be relatively low, the overhead involved in distributing images rather than text is reasonable. Unedited ASCII text could easily be provided along with (or instead of) the images through omnifont text recognition hardware now installed at NAL, however, if this was desirable.

- Although selections can be made for the databases developed by the Text Digitizing Project, giving due consideration to copyright restrictions, it is likely that a substantial part of the interlibrary loan requests received will require the scanning of copyrighted material. In general, it is also desirable to store the scanned material at NAL as part of the overall development of electronic databases. Copyright restrictions may preclude this, however. In addition, indexing may be required if stored images saved as a by-product of responding to an interlibrary loan request are to be fully usable to satisfy future interlibrary loan requests or for inclusion in a database published as part of the digitizing project.

The service improvement to be gained by satisfying interlibrary loan requests electronically is twofold. First, delivery would be faster. Second, since the material would arrive at the requesting library in electronic form, the requesting library may be able to deliver it to the patron in a more convenient form and/or in a more expeditious fashion. Each of these factors requires some scrutiny, however. If the majority of the time involved in responding to an interlibrary loan request is spent in finding the material at NAL rather than copying it and sending it out, then an improvement in delivery speed will make little difference. Improved delivery speed will be noticeable to the user only if interlibrary loan requests are serviced quickly at NAL. At the requesting library, copyright restrictions or lack of an appropriate local infrastructure may preclude quick delivery of the received material to the patron (in either electronic or paper form).

The use of fax as a means of delivery of documents requires specific discussion. There is a growing population of fax machines both in libraries in the U.S. (and internationally) and in the hands of end users of libraries. This ubiquity makes fax technology attractive and worthy of serious consideration as a document delivery channel. Over the past decade a number of experiments have been conducted to explore the use of fax technology to support interlibrary loan activities. These experiments have met with at best limited success, and have revealed a number of serious problems with fax technology, including:

o Quality of fax reproduction. The resolution of a typical fax transmission is low and is often unsatisfactory for material containing type in small point sizes, diagrams, or equations with subscripts or superscripts. In addition, inexpensive (under $10,000) fax machines typically print on special paper rather than plain bond paper, and user reaction to the combination of shiny paper and low resolution is often extremely negative.

o Because fax machines (except for expensive ones) don't store documents for later transmission, serious operational problems have arisen.

o Fax machines (at least The Group III machines currently available) want to operate over the switched (dial) telephone network. (Some machines will function with leased lines, which are not applicable to NAL's situation.) The Group IV machines - still not generally available commercially - will work with a X.25 based packet switched network. Existing machines will not run over the national research internet. This leads to the need to install special phone circuits for the fax system, and very high telecommunications costs.

I would suggest that NAL treat fax as follows for document delivery:

(1) Accept request via fax.

(2) Deliver documents to the existing base of fax machines as an option, but do not consider fax machines as the primary delivery channel for documents. This primary delivery channel should be a new device which was over the national internet (using TCP/IP) and which will accept documents imaged at higher resolution than the standard fax (say 300 dots/inch and above). The existing fax network should be considered a secondary delivery channel, one and that is only capable of handling degraded quality document images in most cases.

(3) As higher quality (i.e. higher resolution, plain paper) fax machines become more common, such machines can be regarded as "quality" document delivery devices., but expensive ones due to the cost of communicating with the machine over the switched telephone network.

2.0 Roles for Telecommunications and Computer Networking

Telecommunications and networking technologies can serve two major and distinct functions: They can provide access to data, or they can serve as a means of distributing data. In the former case, a user employs the network to select documents of interest from a document database that is stored at another site. The selected documents are then delivered to the user across the network. In the latter case, the network is used to transfer a document database automatically from one site to another. The end user searches a copy of the document database without any use of the network employed to transfer it from site to site. Obviously, one can construct systems that combine these two functions in arbitrarily complex ways.

Provision of access across a network requires a unique set of transmissions for each user, since each user will issue his or her own unique requests and will expect a response from the system. Distribution of data, in contrast, is a broadcast-oriented operation in the sense that the same data is transmitted to a large number of receiving sites. The translation of the general idea of broadcasting into specific network and telecommunications technology is quite difficult, however; this will be discussed in section 2.4.

Another important distinction between distribution and provision of access is the level of demand it places on the computing facilities at NAL. For distribution of data, a relatively small computer can be used. All this computer has to do is to transmit data across the network, deal with acknowledgments from receiving sites, and occasionally retransmit data that a receiving site did not get correctly on the first try. To support remote access, NAL would have to install and support locally either a number of single-user workstations connected to the network or a large, multiuser host. Workstations would have to use specialized retrieval software that referenced magnetic or write-once disks rather than the standard distribution CD-ROMs so that they could have access to a changing database.

Under this taxonomy, document delivery in an interlibrary loan context is an access activity rather than a distribution activity; it is essentially a specialized form of query and response serviced from a central site.

2.1 Telecommunications and Networking as a Supplement to Optical Publishing

The basic shortcoming of optical publishing is that a significant cost is associated with each publishing run, and there is a significant delay between the beginning of the publication cycle and the time that the CD-ROM or other distribution media is in the hands of the user community. Thus, for anything other than a static database, the latest CD-ROM distribution is always somewhat out of date, even on the day the end user receives it.

Telecommunications and computer networking can address this problem in two ways: New data can be distributed to sites that received the CD-ROM, essentially allowing each site to maintain a completely up-to-date database, or user queries can be run against the local CD-ROM database and then (possibly at the user's option, and only for queries that really require current data) passed through the network to NAL for evaluation against a small database of information that has been added or changed since the last CD-ROM was issued. In the latter case, the network is providing access to information as opposed to distributing it.

Functionally (leaving economics aside for the moment), each approach has its pros and cons. The advantages of distributing data are that end users not rely on the availability of a remote database to obtain current information, since a current copy of the database is always available locally; the load on the central facility at NAL is predictable and does not depend on user query activity at participant sites; to a great extent, distribution of data can be scheduled at NAL's convenience; and the user does not need to make a judgment about the importance of accessing the master database at NAL to obtain current data. The major disadvantages of distributing data are that the design of the database retrieval software at user sites must provide for repeated incremental updates, including the ability to add access points for newly received material to retrieval indexes (which implies substantial additional system complexity); and the system design must provide for the handling of situations in which a site fails to receive a periodic update or loses an update it has received (through a magnetic disk crash, for example).

Conversely, the advantages of using the network to provide remote access to data that postdates the current CD-ROM are that the user is ensured of always having current data (if some care is taken to communicate the publication date of the current CD-ROM at the user site to the central host as part of the query); there is no need for the system complexity necessary to provide local update capability for systems at user sites; and there is no need to worry about whether remote sites have received updates to the database. The problems with remote access are that the central site at NAL may have to deal with a heavy and somewhat unpredictable query processing load and that the end users throughout the community become dependent on the availability of the central site and the network to obtain up-to-date data. In addition, response time as perceived by the end user may suffer severely if an image-transfer format is used. In data distribution, the transfer of images is a batch process that can take place during off-hours, whereas in the case of remote access the user must actually wait for images to be transferred across the network.

2.2 Telecommunications and Networking as an Alternative to Optical Publishing

It may prove desirable either to eliminate the optical publishing mechanism for database distribution altogether or to allow sites to choose whether they receive data on optical media or over the network. Again, two basic approaches can be used: distribution or remote access.

In the case of remote access, NAL would essentially act as a central database utility for the user community, which would transmit queries to the NAL system for execution and then receive selected documents. The advantages to this approach are that it eliminates all of the complexity of local data distribution, and the database being accessed is always current. The major disadvantages are that the user is absolutely dependent on the NAL computing facility and the network for access to the database and that NAL must run a full-scale online service bureau operation. Also, as is the case with supplementary remote access, one must worry about the length of time it takes to transmit an image across the network to the end user.

If the database were distributed to remote sites, each site would still most likely receive the initial version of this database on some kind of physical medium (tape, CD-ROM, diskettes, and so forth). to each remote site. All subsequent updates would be received over the network. In theory, it would be possible to transmit the initial version of the database over the network, but this would probably be prohibitive in terms of cost and time, particularly when new participants wanted to obtain a large existing database, which would have to be sent over the network all at once. Thus, the distribution case really differs little from the situation already discussed in which supplementary distribution is used.

2.3 Telecommunications, Networking, and Remote Access

If sites are to use remote access to NAL (or to other locations housing copies of the database), there are only two real alternatives: Either a dial-up telephone connection must be established between the remote site and the central facility when queries are to be submitted, or both the remote and central sites must be connected to a computer network (or internetwork). The essence of remote access is bidirectional point-to-point communications.

The various specialized telecommunications technologies discussed later in this report for one-way broadcast distribution of data are thus not applicable to remote access except in the following very limited sense: If a two-way, low-bandwidth network is in place that allows a remote site to establish a connection to a central site in order to submit a query and to acknowledge receipt of a response, the actual response to the query (particularly if it is an image, which requires a high-bandwidth transmission medium) can be tagged for a specific receiving site and sent over the one-way broadcast distribution channel. All remote sites would listen to the broadcast distribution channel and capture data tagged for the site in question.

Remote access to data thus implies not only central facilities (that is, computers and a database) to respond to queries from user sites, but also either a two-way network providing high-speed broadband transmission back to the user site (particularly if images are being transferred), or a low-speed two-way network supplemented by a high-speed one-way data-distribution channel that can be used to send responses back to the originating sites.

2.4 Telecommunications, Networking, and Data Distribution

The data-distribution approach has more complex interactions with telecommunications and networking than does an approach involving remote access. In data distribution, one wants to distribute the same data to many sites--in essence, to perform a data broadcast that all sites can receive.

To accomplish this, one can look directly to broadcast-oriented telecommunications media. These include various satellite-based data-distribution methods, such as the packaging of data in the vertical blanking interval (VBI) of a television signal (proposed by the Public Broadcasting System); direct encoding of digital data into a TV signal (proposed by Howard Hilton); or one-way satellite transmission to small earth stations (so-called very small aperture terminals, or VSATs; these are used in a number of real-time data-distribution applications, such as to distribute stock prices to brokerage houses). The difficulty with the use of such a broadcast medium by itself is that the central site receives no indication that remote locations have received the data successfully. A site could fail to receive a given transmission for any number of reasons, including hardware failures, full disks, receiver problems, and operational errors such as the system being turned off. To make broadcast-oriented telecommunications work reliably for data distribution, one must in effect turn the system into a network by adding a two-way response channel, across which the remote sites would acknowledge successful receipt of a data broadcast. If a reception failed at a given site, the data would be rebroadcast at a later date. Such an acknowledgment can be achieved either over a low-speed network that connects all sites or by a dial-up modem connection over voice telephone lines.

In effect, then, the use of broadcast-oriented media on a practical level requires the same configuration needed in one of the scenarios involving remote access to data: a high-speed one-way broadcast distribution channel and a low-speed (possibly dial-up) two-way communications network between the central transmitting site and the receiving sites.

Note that the use of broadcast-oriented telecommunications to distribute data involves some significant system design issues, even given the existence of a network to provide acknowledgments. For example,

- The central site, after broadcasting an update, can get one of three status indications from each remote site: a successful receipt indication; an unsuccessful receipt indication (for example, that the site was running, at least to some extent, but could not actually capture the broadcast data for some reason); or nothing at all (indicating that the site was completely out of service or was having problems so severe that it did not realize a broadcast took place). When should the central site rebroadcast the data in this last situation, and how many times should it be willing to rebroadcast? What are the effects of having one or more receiving sites out of service for a long period of time, and how are they resynchronized when they resume operation?

- When a receiving site has been out of service for some period of time, how does it know how many broadcasts it has missed and notify the central site that rebroadcast is required? If a receiving site is having severe problems with its telecommunications receiving equipment, how does it know that it has missed a broadcast (as opposed to just thinking that the central site did not broadcast for some days)?

- Situations can arise in which a transmission fills the local storage on receiving sites. In such cases, part of the transmission will be received, but manual intervention at each site might be needed in order for the full transmission to be captured. Similarly, receiving sites will have to be designed in such a way that they can fail in the middle of a transmission without damaging their local databases.

- What do the remote sites think if a major hardware or telecommunications problem occurs at the central site and prevents it from broadcasting data for a few days?

None of these problems are insoluble. They are listed to demonstrate that designing a reliable system based on a broadcast channel and an acknowledgment network is a fairly complex engineering effort, and that a great deal of system complexity will be required to ensure that the system functions in a stable and reliable way.

The other alternative for data distribution is to connect all sites to a common computer network (or internetwork) and to perform "broadcast" data distribution over the computer network. Before addressing this possibility, we must first review some of the protocol and architectural concepts involved in computer networks and internets.

An internet is a network of networks linked together by gateways. Within an internet (which includes the simple case of a single network), hosts can establish bidirectional point-to-point links for reliable data transmission. These links are managed through the use of transport protocols (such as the Transmission Control Protocol [TCP] in the Department of Defense TCP/IP protocol suite, or TP4 in the ISO protocol suite). These transport protocols ensure the reliable exchange of data through checksums, acknowledgments, and message sequencing. They will correct for messages that are garbled, delivered out of order, lost, or duplicated.

An individual network within an internet may be built on a telecommunications medium that inherently supports the sending of a broadcast message (that is, a message sent to every host on the network) or a multicast message (that is, a message sent only to those hosts on the network that have identified themselves as being members of a specific multicast group. Examples are networks based on packet satellite technology, certain types of packet radio networks, and networks based on a common bus such as Ethernets or token ring networks. The internet protocol level (the layer below TCP in the Department of Defense protocol suite, or the upper level of the network layer in the ISO hierarchy, just below the transport protocol) provides facilities for internet broadcast and multicast.

When the underlying network provides broadcast or multicast capabilities, the internet protocol maps the internet broadcast into a specific network broadcast or multicast, where appropriate. In theory, a network that does not inherently support broadcast or multicast messages can simulate them simply by sending a copy of the broadcast or multicast packet to each host that is supposed to receive it as a series of point-to-point transmissions, but for efficiency reasons existing networks do not, by and large, actually perform such simulation.

Internet protocols are unreliable datagram-based protocols; they simply make a best effort to deliver a packet and do not guarantee delivery, correct sequencing, or nonduplication of packets. For point-to-point connections, a transport protocol such as TP4 or TCP performs the functions necessary for reliable data interchange. The definition of protocols to perform reliable data transfer in a broadcast or multicast connection is an active area of research within the computer networking community today; such protocols are not standardized and will not be found in commercial offerings.

Consequently, if NAL wants to use a computer network to perform "broadcast" data distribution, it will have to take a leadership role in this specialized area of computer networking. This is not necessarily bad: such a capability will be important for a wide range of distributed database applications in the coming years, and it seems likely that grant funding could be obtained from some research-support organization (such as NSF or DARPA) to help with the necessary research and development. It must be clearly understood, however, that this is a research and development project.

Several alternatives exist to the actual design, definition, and implementation of a reliable multicast protocol. One would be to define an applications-level protocol that was specialized to NAL's data-distribution application and that ran on top of an internet protocol. Practically speaking, this would amount to defining a special-purpose reliable multicast transport protocol. Hopefully, an application-specific protocol would offer some reduction in design and implementation complexity when compared to a general purpose solution.

A simple approach using only existing off-the-shelf technology would be to have the central site open a series of point-to-point reliable connections (using, for example, TCP or TP4) to receiving sites and transmit the data across the network repetitively. This is inefficient but simple to implement and highly reliable. Interestingly, this is the approach taken by the current crop of commercial distributed database systems; they employ this technique precisely because no standard reliable multicast protocols are currently available upon which the distributed database software can build.

Another alternative would be to distribute data through the network as a series of point-to-point reliable transmissions. One would simply map out a spanning tree for the sites to be updated, and sites would receive and then retransmit the updates through the network. This is preferable to the simpler approach of merely running a series of reliable point-to-point updates through the central site, since it would minimize the amount of data to be passed redundantly through the network. It would be complex to implement, however, since one would have to accommodate the possibility of failure of repeating nodes in the spanning tree. Again, implementing this transfer mechanism would require considerable research and development.

3.0 Technology Issues and Alternatives

This section considers various technologies that must play a role in any network information-delivery system for NAL. Two separate areas are discussed:

- Telecommunications technologies that can be used to link NAL with remote participant sites
- Protocols that will be used to communicate among the sites

3.1 Telecommunications Technologies and Interconnectivity

Data-transmission technology presents a bewildering range of alternatives that might serve as building blocks for a system. To make matters worse, these can be combined together in a number of different ways to produce hybrid systems. The previous discussion indicates that, for example, many possible system configurations require a hybrid consisting of a low-speed two-way network and a wideband data-distribution channel that may be based strictly on telecommunications rather than on networking technology or, alternatively, a highly asymmetric data network that can service queries coming into NAL and return digitized documents to the remote sites in response to these queries. This subsection considers the following building blocks and approaches:

- One-way television-based telecommunications media. This would be used in conjunction with a low-speed two-way network.

- Two-way VSAT (very small-aperture terminal) communications, offering a high-bandwidth link from the central site to the remote sites and a low-bandwidth link from the remote sites back to the central site. This could serve as a complete stand-alone network.

- One-way digital satellite-based telecommunications media. This would be used in conjunction with a low-speed two-way network.

- The national internet, both as a low-bandwidth two-way network to be used in conjunction with a broadcast channel and as a stand-alone bidirectional data-transmission facility in its own right.

- The use of existing networks that serve the library community, such as the OCLC telecommunications network.

- The use of the public switched (dial) telephone system.

- The construction of a special-purpose low-speed two-way network to be used in conjunction with a high-speed one-way broadcast channel.

- The construction of a special-purpose high-speed (asymmetric) bidirectional network, to be used for either data distribution or remote access.

3.1.1 Television-Based Satellite Broadcast Media

Television signals are simply a special form of radio transmission. Radio propagation at the frequencies used for television broadcast is basically on a line of sight between transmitter and receiver. Television signals are transmitted nationally or internationally through the use of satellites.

A satellite used for television contains a number of transponders (typically about 30 for current state-of-the-art satellites). These usually operate in the C band (6Ghz for uplink, 4Ghz for downlink) and are not particularly sensitive to weather interference. Large dishes are required for satellite uplinks; smaller dishes can be used to receive the signal from the satellite.

Each satellite transponder is set to a different frequency within the 6GHz band used for uplink and the 4Ghz band used for downlink. In addition, through polarization, two signals can be passed through each transponder in parallel. A transponder is simply a frequency-shifting radio repeater; it receives a signal in a specific segment of the 4Ghz band used for uplink, translates it into a specific segment of the 6Ghz band used for downlink (with little or no modification of the signal's power spectrum over the band, other than perhaps some filtering), and retransmits it though an antenna to a very broad geographical region that is covered by the satellite.

A typical satellite will cover the entire continental United States and perhaps some parts of the Caribbean, northern Mexico, and southern Canada, and will often use a special directional antenna (called a "spot beam") to extend the signal to Hawaii and/or Alaska. Other parts of the world require multiple satellite hops (from earth to the satellite, back to earth, and then up to another satellite). Direct satellite-to-satellite links are in use, at least experimentally, by NASA and the military but are not generally available for commercial use (although they may be with the next generation of satellites).

Virtually all television stations (and many cable TV operators) are now equipped with satellite receiving facilities; they use this facility to receive programs originated by the networks. The networks, and a few large individual stations that originate programming (including some educational institutions that originate educational television programs) have their own uplink dishes that they use to send programs up to a satellite transponder. Although a satellite transponder is not normally specialized to a television signal (as opposed to some other type of radio waveform), the uplink and downlink equipment used at most television stations is specialized to the TV waveform, in terms of both modulation and bandwidth of RF.

The standard NTSC TV signal used in the United States contains 525 interlaced scan lines that form a picture. In addition, there is some dead space in the signal called the vertical blanking interval (VBI), used to allow the cathode ray gun that generates a picture on a television set to return from the bottom of the screen back to the top. This VBI information is ignored by a normal TV set. It is used for a number of applications, including teletext, in Europe and (to a very limited extent) in the United States.

PBS Enterprises, a commercial subsidiary of the national Public Broadcasting Service, proposes to use the VBI as a one-way data-communications channel, both in specific areas (for terrestrial television) and nationally. The drawbacks to this approach are that

- Special hardware is needed to receive and decode the signal either from local rebroadcast or from a satellite downlink. In addition, if a satellite downlink is used, a TVRO (TV receive-only) antenna and related electronics are required.

- Substantial coordination is needed between the PBS central site and the local rebroadcasters to ensure that the VBI is propagated into the necessary geographic areas; PBS would (at least in theory) take care of this, however.

- Scheduling is a major concern. Local stations often take a network feed and store it on videotape for later rebroadcast; it would be necessary to schedule retransmission with each local broadcaster.

- The service is expensive.

- The service is quite low bandwidth (less than 10Kbits per second).

In addition, because of the need to schedule rebroadcasts with this service, it could not feasibly be used as a real-time response channel except at extraordinary cost. (Doing so would involve placing a real-time link between NAL and the PBS uplink in the Washington, D.C., area and licensing use of the channel for large parts of the day or even on a dedicated continuous basis. Even then, it is not at all clear that the local rebroadcasting affiliates would propagate the signal downlinked from the satellite in real time.)

Howard Hilton has proposed a related scheme in which data is actually encoded in the picture segment of the NTSC signal. Devices are available from such companies as Kirsch Technologies to decode data from the TV signal. Hilton's scheme has the advantage of supporting a much higher data-transfer rate. However, it suffers from the same scheduling problems as the VBI solution, with the added drawback that a whole television channel would have to be dedicated to the transmission. This channel could be received directly from the satellite through the use of TVRO technology or could be rebroadcast locally through special arrangements with local television broadcasters in each area NAL needed to cover.

Operationally, Hilton's scheme is also rather cumbersome. Each receiving site would normally connect a videocassette recorder (VCR) to a television receiver (either a TVRO downlink or a standard broadcast) and capture the TV transmission containing the encoded digital information. This would require that each receiving site know the transmission schedule and have the local VCR ready. At some point after the recording was made, it would provide data to some sort of database update program running on the local site's PC or workstation, which would use the VCR with a special decoder board as an input device. The received data would have to be either moved to local write-once or magnetic storage on the workstation or left on the videocassette. In the latter case, it would still be necessary to maintain indexes to the data stored on the videocassette in volatile storage on the workstation and for the local site to manage multiple videocassettes.

In my opinion, these technologies are not appropriate for the NAL project. They are too expensive and unmanageable. If a satellite data-distribution mechanism is to be used, VSAT technology (discussed next) is much simpler to manage and certainly no more expensive.

3.1.2 Non-Television-Based Satellite Technologies

Other types of satellite systems encode digital data directly encoded into a radio signal with one of several modulation techniques, instead of trying to make the digital information look like a television signal. Because the RF techniques involved are specifically designed to encode digital data, they are much more efficient than encoding schemes based on television signals.

Satellite networks oriented toward the transmission of digital data typically operate on either the C (6Ghz/4 Ghz) or K_u (14Ghz/12 Ghz) bands; the C band is older technology. The advantage of the higher-frequency K_u band is that a smaller antenna can be used due to the shorter wavelength of the signal. The disadvantage of this band is that it can be vulnerable to weather; heavy precipitation (rain or snow) can cause signal scattering and reception failures. A number of approaches, including variable power uplinks and transponders, are used to circumvent this weather sensitivity. (Although it should not happen in theory, in practice C band transmission is also sensitive, though to a much lesser degree, to severe storms, in my experience.)

NASA is also planning to launch a satellite called ACTS (Advanced Communications Technology Satellite) in the early 1990s that will operate on the K_a band (at still higher frequencies) using a very elaborate multiplexed, scanning spot-beam system. They are actively seeking collaborators, and NAL might wish to examine the possibilities of participating in this program if it wishes to pursue small-antenna satellite communications.

Three years ago, it appeared that the industry was headed for a satellite transponder glut by the end of the 1980s due to the large number of communications satellites scheduled for launch. However, with the Challenger disaster and the subsequent delays in the STS program, it appears that there may in fact be a severe shortage of transponder capacity in the early 1990s before expendable-launch-vehicle technology becomes generally available and compensates for the large launch backlog. The lifespan of a communications satellite is normally governed by the amount of hydrazine that the satellite carries for station-keeping maneuvers. Once this propellant is exhausted, the satellite is moved into a final resting place outside of geosynchronous orbit and retired. The hydrazine supply in a typical communications satellite lasts 5 to 10 years; many of the existing satellites will exhaust their hydrazine supply by the early 1990s. This will drive the cost of satellite communications upward.

There are two basic digital satellite-transmission architectures. The first is a standard series of uplink/downlink dishes running either in C or K_u band and functioning as a private satellite terminal network. Virtually all systems currently available from commercial vendors use frequency division multiplexing, time division multiplexing, or a combination of the two to share the frequency bandwidth among multiple transmitting stations. This architecture is likely to be impractical for NAL for several reasons: The NAL site network will be too large for this type of approach; the cost of civil engineering for dish installation will be prohibitive; and the costs of the hardware and transponder time will be prohibitive. The University of California's experience, for example, is that $200,000 is required to purchase and install equipment for a typical site.

The second major architectural approach for digital satellite communications is the use of small earth stations (VSATs). One of the most limited resources in a satellite communications system is power; as a rule, a powerful uplink run through a large dish can be received by a group of smaller dishes. In addition, the smaller dishes can transmit lower-bandwidth uplinks that can be received by the large central dish that serves as the main broadcast source in the system.

VSAT systems are used today by a number of major corporations that need to distribute real-time data--stock brokerage houses distributing stock prices, retail chains such as K-mart distributing inventory information, and the U.S. Forest service. Typically, the antennas at the remote sites are less than 2 meters in diameter. The ability to use these small antennas is achieved at the cost of a very large (10- to 12-meter) central site ("hub") antenna. The hub antenna can be either dedicated and installed at the central site of the VSAT system user or shared and installed at the central site of the VSAT system vendor, in which case each user organization runs a communications link between its central site and the VSAT hub.

In general, communication in a VSAT network occurs between the remote sites and the hub; a message from one remote site to another is forwarded through the hub, adding about a half-second propagation delay. The sharing of the frequency among the remote VSAT stations is typically handled through polling from the hub.

VSAT networks, by their design, offer asymmetric bandwidth. A typical remote downlink might "hear" a shared channel running at 1 million bits per second or more, but will have an uplink running at only 9.6Kbits to 56Kbits per second.

Functionally, VSATs are a good match for the needs of NAL if it wishes to pursue either distribution or remote access. VSAT technology is not inexpensive, however. A typical site can be installed for $10,000 to $15,000, including civil works, which are minimal with the small antenna. Often, the most expensive part of the entire installation is the cabling to run data from the rooftop or the ground outside of a building to the site at which the computing equipment connected to the VSAT is actually installed. The major expenses are the cost to purchase or lease time on the hub station and the transponder time. In addition, some delicate interface issues between the VSAT stations and the workstations at the remote sites would have to be resolved.

3.1.3 One-Way Satellite Transmission

Either the standard C or K_u band earth-station network or the VSAT network configuration can be run in a one-way mode. This allows smaller antennas in the non-VSAT configuration and much simpler node hardware configurations at the receive-only sites. Oddly enough, however, this configuration is not offered as an off-the-shelf item except in VSAT configurations; in a VSAT system it simplifies both the licensing (no site clearance through the FCC is needed for a receive-only station; for a normal two-way network a time-consuming and expensive frequency coordination study is required for each transmitting site) and the hardware configuration on the remote VSAT sites.

This system configuration is actually quite similar to Hilton's scheme when used with TVRO facilities. The major differences are that with a VSAT or radio signal the complexities of encoding data within a television signal are eliminated; the satellite channel can be sized to the needs of the system, rather than being fixed at the size used to carry a TV channel; data received from the satellite would be captured and processed in real time by the local workstation at each receiving site; and the satellite terminals would likely be more expensive, since they would consist of specialized digital equipment rather than mass-produced TVRO configurations augmented by a decoding interface board. The extra capital cost of this receiving system would probably be quickly offset by the lowered cost for satellite transmission facilities, however, since this configuration would use satellite transponder capacity more efficiently.

In many ways, a special-purpose one-way satellite system using small receive antennas is very attractive as the backbone of a data-distribution network. FCC licensing would not be necessary at the receive sites, and hardware costs could be kept reasonably low (probably under $10,000 per site). Site installation costs would be highly variable and would depend to a great extent on the difficulties of running cable between the satellite dish and the location housing the receiver electronics and workstation. By using packetized transmission with a mix of multicast and individual site addresses, the system could also serve as the wideband half of a data-access system for query response and demand document delivery.

Such a one-way system would have to be supplemented by a narrow-band system to allow the remote sites to pass requests and receipt-of-data acknowledgments back to NAL, and some creative protocol engineering would be required to make the whole system work. NAL would need a link from the local system at the library to a satellite uplink feeding the system. To save transponder costs, it might be possible to arrange for transponder time only during limited hours (perhaps at night). Finally, if the remote site systems were properly designed, it might be possible at some point in the future evolution of the system to upgrade at least selected remote sites to allow two-way satellite communications as traffic from the sites to NAL or to other hub sites grew to a level that justified such communication.

A final point should be made about all satellite-based data-communications systems, however: In my experience they need a substantial amount of ongoing maintenance, especially at transmitting sites. Transmitters are governed by FCC regulations and require regular monitoring; this service can sometimes be purchased from the system vendor, although at a nontrieval cost. Uplink transmitters may suffer from frequency drift or other problems that require on-site adjustments or component replacements.

When initially installed, each dish needs to be properly pointed at the satellite being used by the system; over time, the system may have to migrate from one satellite to another, requiring physical reorientation of each dish that is part of the system and a very complex cutover period when some dishes point to the old satellite and some point to the new satellite. A shift from one satellite to another may also require frequency changes at the earth stations, requiring potentially costly hardware changes. Further, spares may be a problem since each transmitter is normally configured for a system-specific frequency, and spare transmitter components would be maintained only if NAL paid for them to be manufactured.

Finally, satellite systems are vulnerable to brief outages at predictable times when the sun is directly behind the satellite, interfering with the radio signal. While these are brief (two hours at most), predictable, and not terribly frequent, they are a nuisance.

3.1.4 The National Research Internet

In the early 1970s, the ARPANET linked a number of major universities, defense contractors, military and government installations, research labs, and similar installations together into a computer network. This subsequently evolved into the DARPA internet, a network of networks that linked several wide-area networks (the original ARPANET, now split into ARPANET, a research network, and MILNET, an operational military network); local area networks at many of the participating institutions; and other wide-area national and international networks.

In the mid-1980s, the National Science Foundation (NSF) began a major interinstitutional computer networking project, NSFNET, to link universities to the NSF supercomputer centers located around the country and also to interconnect the supercomputer centers themselves through very high-speed trunks. NSF subsidized the establishment of a number of regional networks in areas such as New York State (NYSERNET) and northern California (BARRNET). Over time, the interconnections between the NSFNET and the DARPA internet multiplied and the two networking projects dovetailed, with explicit cooperation between the two agencies to build a national research internet.

The national internet concept continues to gain momentum. In June 1986, Congress passed legislation introduced by Senator Gore that charged the Office of Science and Technology policy with the task of conducting a networking study. An interagency group called the Computer Network Study Planning Group was established under the auspices of the Federal Coordinating Council for Science, Engineering, and Technology (FCCSET). This group included participants from the Department of Defense (including DARPA), the Department of Energy, the Department of Health and Human Services, NASA, the National Bureau of Standards, and the National Science Foundation. (As far as I can tell, the Department of Agriculture was not part of this effort.) The Computer Network Study Planning Group delivered a report in August 1987 calling for the establishment of a very high-speed national research internet. The NSFNET can be considered the beginning of the implementation of this plan; however, long-term funding for the NSFNET effort, particularly for the upgrades to very high-speed trunks, is clouded at this time.

Until recently the NSF/DARPA network was based primarily on 56Kbit leased telephone lines, and parts of the network were severely overloaded. NSF is moving toward T1 (1.544Mbits/second) trunks with even higher speed (ultimately T3, or 45Mbits/second) trunks on the backbone and linking the supercomputer centers. The NSFNET T1 backbone became operational in July 1988. With the success of NSFNET, DARPA is rapidly phasing out the research ARPANET. ARPANET sites will migrate to NSFNET, and ARPANET will ultimately be replaced by a new very high speed Defense Research Internet, which will continue DARPA's commitment to networking research.

Our survey of the participant sites indicates that we can expect between 25 and 50 percent of the sites to be connected to the national internet. One approach that might be considered is some form of subsidy to expedite connection of the remaining sites to the internet. This would offer a very large payoff in terms of overall network resource accessibility, and some form of joint funding agreement might be worked out with NSF, DARPA, or other agencies. If the Department of Agriculture can become a participant in the National Research Network effort and if Congress makes funding available for the project on a broad basis, it is possible that NAL, Cooperative Extension sites, and the participating land-grant institutions and other organizations identified in Director Howard's plan for a National Agriculture Library and Information Network might be connected to the national research internet as part of such congressional funding.

To become connected to the national internet, a site needs to be TCP/IP capable and to install a link to the nearest site that is already participating in the internet. Note that as of early 1988, NSF awarded a contract for the management of the NSFNET to a consortium consisting of MERIT (the University of Michigan), IBM, and MCI. This may radically change the rules and parameters, but the effect of this new management structure remains to be evaluated.

One side issue: A high priority should be to connect the central site at NAL to the national internet. This would allow immediate use of the internet both for remote access and for data distribution to user community sites that are already a part of the internet and would also provide an incentive for non-internet sites to obtain a connection to the internet.

The funding basis for the internet is somewhat unusual. There is no traffic-based charge for the use of the internet; instead, trunks and packet switches are funded directly by NSF, DARPA, and participant institutions as overhead. Thus (as I understand it), if NAL obtained an internet connection, it would not have to pay to use the internet to communicate with other internet sites. One additional attractive by-product of the internet connection is that NAL would get immediate connectivity with a large number of other institutions in the United States that are not part of the Text Digitizing Project but that might be interested in it and ultimately join the project.

3.1.5 A Special-Purpose Network

On both technical and economic grounds, a case can be made for a dedicated network linking NAL and the participant sites through an asymmetric mix of satellite trunks and low-speed dedicated or switched land lines or connections to the national internet. The argument in favor if this approach is that specialized technology and system architecture would be matched to the needs of the NAL document delivery effort for optimum performance and minimum cost. Although it may be that the investment necessary for such a fully custom system cannot be justified given NAL's current priorities and objectives, the specialized nature of the system at least gives rise to some clear advantages.

Another alternative would be simply to construct a traditional packet-switched network that linked the participant sites and NAL using standard symmetric telecommunications trunks. I do not believe that this alternative makes sense for the following reasons, and consequently I have not examined it in detail:

- Such a network would be very expensive. If an investment on this scale is to be made, it should be for a system that is tailored to NAL's unique document distribution and transmission requirements. The counterargument to this would be that traditional packet-switched networking technology is well understood, whereas a custom asymmetric system is unique and consequently represents a somewhat risky research and development project. I feel that these risks could be managed, although construction of such a custom network would unquestionably be a major undertaking requiring careful management.

- Such a private traditional packet-switched network duplicates the existing and planned National Research Network initiatives. Unless the Department of Agriculture is precluded for political reasons from doing so, there are major financial, technical, and management advantages in participating in the national research internet rather than setting up a parallel network. The existing national network initiatives are of sufficient scale to allow the incorporation of a high degree of redundancy; to attract the best networking people in the country as participants in the planning and running of the network; to be able to afford very advanced technology and support custom technology development where needed; and to reach many sites that would make some use of NAL's services but that could not justify a dedicated connection to an NAL network.

3.1.6 Existing Library Telecommunications Networks

Currently, most libraries (including most of the participants in the Text Digitizing Project) are linked to a network run by OCLC, Inc. of Dublin, Ohio, that supports shared cataloging. Each library has one or more terminals connected to the central system in Dublin through this network, which is sized to carry interactive terminal traffic. The network is almost entirely idle at night, since the terminals are used by catalogers during library working hours. One of the most attractive aspects of this network is that it represents a resource that is already in place and financed and that already services the great majority of sites with which NAL wishes to communicate. (And, for those sites that are not already connected, connection would offer multiple benefits if NAL used this network, since the site would obtain access to both NAL services and OCLC services. This would help to justify the expense of connecting to the network.)

The OCLC network is privately owned, of course, and NAL would have to work out an arrangement with OCLC if it wished to use this network. The costs of such an arrangement are unclear. There are several other drawbacks to making a commitment to using the OCLC network, even if such an agreement could be successfully negotiated:

- The OCLC network is in a considerable state of flux. OCLC has for a number of years been working on a massive system conversion effort designed to move its network from 1960s-style private protocol multidrop architecture to standard commercial X.25. This conversion has encountered major problems and reversals due to the inability of OCLC's vendor to meet its performance specifications.

- The OCLC network is relatively low speed, since it is designed for interactive terminal traffic. It could support some level of digitized image transfer, but it does not appear that it could handle image transfers at a reasonable level during the daytime, and it would be slow even during off-hours.

- OCLC has its own plans to use its network for document delivery and has several active projects (such as EIDOS) to exploit idle network capacity for this function. If these plans are successful, the amount of available capacity on the network may be very limited, even at night. Further, OCLC may not be interested in making its facilities available for a competing document delivery project.

- OCLC's planned network is based on X.25. This is completely inconsistent with the networking approach currently in use at most universities and throughout the U.S. research community. It is possible to incorporate an X.25 network into the internet by using well-established technology (TCP/IP over X.25), but this requires the presence of IP gateways and trunks that currently do not exist at most OCLC sites. In most installations, OCLC's network is a specialized service that is delivered to dedicated cataloging terminals in libraries and is not interconnected with other local network facilities. The cost of such interconnection is substantial ($20,000 or more), and this capital investment would be a necessary prerequisite to any use of existing local area networking at participant sites.

The Research Libraries Information Network (RLIN) of the Research Libraries Group (RLG), based in Palo Alto, California, is also deploying an X.25 network to connect its cataloging terminals at libraries across the country with its central site in California. The RLIN network is somewhat more stable than OCLC's network and might also be considered as a delivery highway. Because of the use of X.25 rather than TCP/IP, RLIN's network raises the same compatibility considerations as OCLC's. In addition, RLIN's network is less attractive as a delivery path because it reaches a much smaller number of libraries. To the best of my knowledge, RLIN is not currently considering the use of its network for document delivery.

3.1.7 The Public Switched Telephone Network

Yet another alternative would be to avoid standard computer networking technology completely and to rely entirely on dial-up telephone lines. This could be done through the inclusion of autodial modems in the workstations. The major cost in this approach is common-carrier charges. There would also be a modest cost to provide a phone line at each workstation. The big advantage of the public telephone network is that a connection need be established only when it is needed. If remote access is very light or data distribution does not occur often or in large volume, this might be cost-effective.

Depending on the funding basis that NAL wishes to incorporate, several methods could be used to move the cost of dial-up access away from the end-user sites and to NAL. For example, NAL could get a dial-up 800 number that the remote sites could use without charge to submit queries. NAL could then install an outbound WATS line that it could use to dial up the remote sites with query responses at a fixed charge. One alternative that should be investigated if an off-hours data-distribution scenario is pursued is the possible existence of in-place 800 numbers and outbound WATS lines at NAL that could be put to use during off-hours at a low cost.

The major disadvantage of the dial-up method, other than the possibility of highly variable costs unless 800 numbers and WATS lines are used, is the very low bandwidth available via the switched telephone network; realistically, the maximum transmission speed we could hope for would be only 2400 to 4800 bits per second, even with very expensive modems. This would not be satisfactory for transfer of images, except perhaps on an overnight basis. Even overnight, it might be necessary to use multiple outbound WATS lines at NAL for transmission in order to respond to all sites that had submitted queries within the past 24 hours.

3.1.8 Regional Networks

To complete the discussion of telecommunications alternatives, it should be pointed out that a multiplicity of technologies can be applied. One potentially cost-effective approach would be to develop a backbone network of perhaps five to ten sites across the country connected by some type of satellite transmission. Other participant sites could be linked to the nearest backbone site by any one of a number of methods: the national research internet, private leased lines, existing links supporting regional consortia, the OCLC network, or the public switched telephone network.

Such a hierarchical approach could make particularly good sense in the context of Director Howard's proposal for a National Agriculture Library and Information Network if regional centers were also key resource centers in the logical network and if regional centers were directly interlinked as well as being connected to NAL. The links among the regional centers could serve multiple functions, providing backup to the links between the regional centers and NAL through alternate path routing as well as facilitating direct communication between the regional centers.

3.2 Protocol Considerations

Today, the DARPA TCP/IP protocol suite is a de facto standard within the university and research communities in the United States and is coming into wide use internationally. Although TCP/IP is a U.S. Department of Defense military standard and is used to some extent within NATO as well, it has no real standing as an international standard. Current activities with the NSFNET and the DARPA internet are all based on TCP/IP. It is a mature, well-developed, and proven protocol suite that is widely supported by commercial hardware and software offerings. If NAL wants to maintain compatibility with the activities of the U.S. community and the NSFNET, it must operate in a TCP/IP environment, at least in the near term.

National and international standards bodies (including ANSI, ISO, and CCITT) have been developing a new set of networking standards (the OSI protocol suite) since the late 1970s. These protocols are intended as national and international standards, but they are not yet completely defined and are not yet well supported commercially. In the long run (say by the mid 1990s), it is likely that the OSI protocols will become the dominant ones for computer networking. As yet, plans for a transition from TCP/IP to OSI are not well defined for the research internet, although there is a commitment to migrate to OSI protocols in the long term.

This state of affairs is important to NAL for at least four reasons:

- It implies that NAL will have to deal with a transition from TCP/IP to OSI over the next five to seven years, which will cause considerable complications in any operational computer network that NAL deploys.

- The U.S. library community has on the one hand almost totally ignored developments in computer networking but on the other hand has been one of the leaders in the adoption of OSI. Since NAL is interested in actually deploying a network, it must deal with TCP/IP in the near term. This is likely to cause some political complications and technical difficulties, particularly given NAL's status as a national library and the Library of Congress's strong commitment to OSI technology.

- Certain applications-layer protocols of great interest to NAL, such as Z39.50, the information-retrieval protocol recently standardized by NISO/ANSI, are defined to operate within the OSI environment. NAL, along with other institutions that have immediate needs for large-scale operational networks, will have to deal with the problems of using these protocols within a TCP/IP environment in the short term and later migrating them back to an OSI environment.

- Strong support for the OSI protocol suite exists within the federal government. A number of government agencies participated in the development of a specification called GOSIP, which defines the use of OSI protocols within the government and calls for the mandatory use of OSI in most cases in new systems within the next two years. There are exemptions for research systems, but it is not clear exactly what the scope of these exemptions is. The Department of Agriculture participated in the development of the GOSIP specification and signed the GOSIP document. Although the NSFNET also has a long-term commitment to OSI, my sense is that they are thinking in terms of a five-year time frame rather than a two-year one. This again may be a source of some political sensitivity.

Practically speaking, the transition from TCP/IP to OSI is a technical problem that will have to be managed. From a functional perspective, once OSI is fully defined, stabilized, and commercially implemented on a broad basis, there should be no real problem in converting from TCP/IP. I recommend that NAL follow the NSFNET community in converting from TCP/IP to OSI as a matter of broad strategy. This would ensure that the technical problems were resolved and that maximum compatibility with NAL's user community was maintained. In the interim, the political problems related to this decision would have to be dealt with.

4.0 Local Redistribution and Access--the LAN environment

Increasingly, we are seeing the development of local area networks at institutions within the user community. In some cases these are single LANs; in others they are actually internetworks of LANs, often with links to the national internet as well. Today, at least at academic institutions, most of these networks are TCP/IP based; within a few years, however, many of them may make the transition to the OSI protocol suite.

There is a growing need to make the databases distributed by NAL on optical media function within these local network environments. Some of the major factors driving this demand area include the following:

- The desire of end users to access NAL-provided databases from office, home, lab, and other remote sites

- A need to provide access to NAL-published databases from multiple, geographically scattered sites

- A need to integrate data from the NAL databases with other parts of the computing environment, including word processing, typesetting, personal database managers, and so forth

- A need to obtain access to multiple databases from a single workstation using a common user interface.

I believe that it will be necessary to accompany published NAL databases with a set of database support software that can function in this environment and that can meet these developing needs. Placing a network interface on the machine supporting these CD-ROMs to allow it to attach to Ethernets and other popular LANs is not much of a problem. IBM PC-compatible software implementations of TCP/IP that can communicate with most of the popular LAN technologies are readily available from multiple sources (MIT, FTP Software, Spartacus/Fibronics, and Network Research Corporation/Fusion, among others). For more sophisticated UNIX-based workstations (such as SUNs, IBM RTs, and DEC MicroVaxes), TCP/IP support is normally an integral part of the UNIX operating system. The major design issue in supporting the use of the databases published by the Text Digitizing Project on local area networks is the applications-level interface. Three approaches can be taken; each has benefits and drawbacks. It could be argued that it will be necessary to implement multiple approaches. The possibilities are as follows:

- A line-by-line terminal server using TELNET. This would allow character-oriented line-by-line terminals to log onto the workstation and, using the native retrieval language of the workstation (which would have to be oriented to line-by-line ASCII terminals), conduct searches and retrieve materials. In theory, either ASCII text or images could be obtained; one could use a file transfer protocol to move the text or images to the user's workstation.

- A database server model. Here the server would implement a database access protocol such as NISO Z39.50 ("linked systems protocol"). The client workstation would run a user interface that supported the same protocol, which would be used to move text or images from the server to the workstation. More important, the user interface running on the user's workstation would be to a great extent database independent, supporting a canonical search language mapped to Z39.50. An alternative (but less desirable) protocol would be the remote database access protocol under development by ANSI X3H2. The query facilities available in this protocol are determined by the query language SQL and are not well suited to textual databases such as those distributed by NAL. One drawback to the database server approach, at least today, is that no client implementation of Z39.50 exists for workstations, and Z39.50 over TCP/IP is still highly experimental. Several institutions, including the University of California, are working with this approach at present, but it is still in the research and development phase.

Note that some modest extensions to Z39.50 may be needed to make the protocol work well for image databases. Also, Z39.50 assumes the existence of other, external standards that can be used to provide data-transfer formats between hosts communicating using the information-retrieval protocol. For bibliographic data, these are the MARC record standards. New working standards would have to be identified (or perhaps defined) for image and ASCII-format document data transfer.

- A network graphics terminal model. If a highly sophisticated graphics-oriented user interface is required, or if images are to be displayed, a third possibility is to write software in the NAL network server machine that acts as a client to one of the network windowing packages such as MIT X Windows or SUN NeWS. This approach has the advantage that it can work with most popular workstations; the disadvantage is that it is fairly complex to program.

The need to support these types of access to CD-ROM databases distributed by the Text Digitizing Project gives rise to a number of standards issues that need to be incorporated into the ongoing publishing efforts of NATDP; these considerations are discussed in section 6.

5.0 Demand Document Delivery

Demand document delivery has a good deal in common with the database access approach described in section 2 of this report. When a remote library requests a copy of a document, NAL locates the document in its collection, converts it to a series of digital images, and transmits these images to the requesting library. However, there are some important differences as well:

- Unlike a query against an existing database housed at NAL, which is resolved entirely by computer, delivery of a document on demand involves considerable human intervention at NAL. The request may not be completely specific, and human judgment may be necessary to determine the exact material being requested, unless the request is for some fully described item located through a bibliographic database, and the full citation or record ID in the database is passed to NAL as part of the request. After the document is identified, it must be scheduled for scanning, scanned, verified, and then transmitted. There is likely to be a considerable time lag (ranging from hours to days) between the arrival of the request at NAL and the transmission of the document back to the requesting site.

- Unless an automated requesting system is used at the remote sites to initiate requests and the remote system's request ID is propagated back to the remote system when the document is transmitted, further human intervention is needed at the requesting site to match documents received with outstanding requests and to forward the received document (in either electronic or paper form) to the requesting end user at the remote site.

- At least until the accuracy of the omnifont character-recognition technology in use by the Text Digitizing Project improves, it is likely that virtually all documents will be transmitted in image form.

Given these differences, any network system that can deliver images to remote sites can also be used to support demand document delivery. Aside from questions of network capacity (which, in turn, are a function of the number of pages delivered per day), most of the additions needed to support document delivery will be to the computers at NAL and the receiving sites. Specifically,

- The NAL site will have to be able to collect from a local scanner images tagged with a destination and a request ID, buffer these images (which will require considerable local magnetic or write-once optical storage), and schedule them for transmission to the requesting site. Similar systems may make sense at other major resource libraries that are net lenders in interlibrary loan traffic.

- Local receiving sites will need to be able to receive such sets of images, match them with outstanding document requests, and either print the images on a local laser printer or forward them electronically across a local area network to an end user's workstation. The local receiving sites will thus also need considerable local disk storage (either magnetic or write-once optical) to buffer received documents.

Many libraries may wish to participate in demand document delivery even if they are not participating in the Text Digitizing Project. Document requests can be submitted to NAL through electronic mail, dial-up telephone, voice telephone, FAX, or even U.S. mail. It is likely that many of these sites will not have a workstation set up with a laser printer, a network connection, and appropriate software to receive documents in the same fashion as Text Digitizing Project participants. Instead, many of these sites are likely to have FAX machines that are already used to receive documents from other sources and that are connected to dial-up lines. To serve this population, it is recommended that a document delivery system at NAL include an option whereby a request can be returned to a receiving FAX address. Instead of scheduling images for network transmission, the NAL system would dial up these receiving FAX sites (probably late at night when the phone rates are low) and emulate a transmitting FAX machine. The quality of service available to these sites will, however, be limited by the receiving FAX machine's resolution.

6.0 Recommendations and Issues

This final section deals with the next steps that NAL might take in exploiting telecommunications and networking technology. It is divided into four subsections:

- Policy decisions. This report has identified a number of alternative approaches that can be taken to networking, depending on the specific goals, priorities, and resources of NAL management. Obviously, not all of these possibilities can or should be pursued, and choices must be made among them. Strategic decisions that need to be made before the next major steps are taken are identified here, along with additional policy issues that must be considered if certain decisions are made.

- Management initiatives. This report has shown that networking in the United States today is complex and in a state of rapid change; it involves many organizations working together. Along with the more technical steps involved in actually implementing networking technology, NAL must become more deeply involved in the issues at an institutional level in order to ensure that its needs are properly considered as part of the national network-planning process and the development of standards that allow communication and data transfer among computers.

- Specific projects. This section identifies several specific technical projects that I believe should be seriously considered for funding and implementation to provide operating experience with networking technology in support of NAL's mission.

- Standards and guidelines for existing projects. Even in an environment that uses stand-alone workstations, the publishing work of the Text Digitizing Project is beginning to encounter compatibility problems created by the evolution of hardware, lack of standards for hardware and software, and changing technology. As we begin to consider the addition of networking capabilities to the database access software that supports CD-ROMs published by the project, new standards issues arise that need to be considered in future software and hardware selection. These considerations have a positive side, however, in that many of the existing standards and compatibility problems can be subsumed by the more rigorous demands of standards for operating in a network environment.

6.1 Policy Decisions

- Does NAL want to supplement its current optical publishing program with telecommunications or network-based database distribution as defined in section 2?

My sense is that there is no clearly defined need to do this and that the high cost cannot be justified for the types of databases currently under consideration by the Text Digitizing Project. Optical publishing is extremely economical and a good match for the needs of the project.

- Does NAL want to provide remote access through telecommunications or network facilities to databases that are mounted at NAL?

My feeling is that this may make more sense but will be quite expensive. I suggest that it be considered only in cases in which optical publishing is impractical; for advanced access to databases being prepared for publication on optical media; and for those sites that require only very rare access to the material in question and already have network paths to NAL through the internet. (The issue of NAL connection to the internet is discussed later.)

- Does NAL want to attempt electronic demand document delivery as an extension of its existing interlibrary loan activities?

To answer this, I think that NAL must first determine how much of the interlibrary loan response time would be eliminated by electronic transmission and how much of it consists of the time needed to locate and copy material at NAL (and consequently unaffected by electronic transmission of the material). After this breakdown is established, NAL should discuss the need for electronic document delivery with its user community to obtain additional guidance.

- Does NAL want to operate large-scale private telecommunications links (based, most likely, on satellite technology)?

This type of facility could be a valuable resource, particularly if NAL makes a commitment to remote access, network-based database distribution, and/or demand document delivery. Such a resource could also serve the broader needs of NAL, the national agriculture information community, and other programs and organizations within the Department of Agriculture. The development of this resource would be a major undertaking, however, demanding substantial time, money, and management. While precise costs cannot be determined without a more specific system definition (which in turn depends on other issues raised in this report), a ballpark figure of $500,000 for initial system development and deployment to between five and ten sites is not unreasonable, and operating costs once the system was installed could be expected to run well over $100,000 per year (mostly for transponder time).

- Does NAL want to make a commitment to support the evolving local network environments in its user institutions and potential user institutions?

I believe that the answer to this question must be positive.

6.1.1 Issues Arising If a Network Is Created

Many networks, particularly those providing service to new user communities for which capacity planning data was unavailable, have encountered difficulties in governance, funding, and the management of growth. These issues must be considered very carefully as part of the planning for any new network. In particular, the deployment and funding model that has NAL paying for some backbone facilities and individual sites funding and installing their own local connections to the network backbone facilities is appealing: It is simple and permits rapid network growth. As the network grows, however, the network manager (NAL) needs control over local connections in order to manage, debug, and tune the network.

In addition, some funding mechanism needs to be established to allow the continued growth of the backbone network as additional sites are added and traffic increases. Ideally, this funding would be arranged in such a way that capacity available on the backbone can always be kept slightly in excess of immediate demand, ensuring good performance. Finally, experience is increasingly showing that large networks need to be managed by some organization with the people and commitment to do the job. Ad hoc management by a consortium of interested parties, while it may be suitable for broad policy formulation, does not work well for ensuring day-to-day reliable operation of the network.

6.2 Management Initiatives

NAL, particularly in the context of Director Howard's proposed National Agriculture Library and Information Network, is poised to become a serious force in national networking, based on its existing leadership in digital document handling (as developed through the Text Digitizing Project), the scope of its user community, and the importance of agriculture to the United States. The basic recommendation here is that NAL must ensure that it is represented in technical, management, and policy discussions about networking that are taking place on a national level. Specifically,

- NAL (or perhaps the Department of Agriculture) should attempt to become involved in the planning for the National Research Network currently being discussed in Congress and developed under NSF leadership and should become a participant in this network.

Similarly, NAL should encourage participants in the Text Digitizing Project and members of the proposed National Agriculture Library and Information Network to participate in these efforts and should take an active role in informing its community of developments and opportunities in this area.

- NAL should attempt to identify and define the networking needs of other Department of Agriculture activities, such as the Cooperative Extension program and to determine how they fit with NAL's needs. Some thought should be given to developing a document that describes broad-based networking needs for Department of Agriculture programs. This would be advantageous in that it would offer the possibilities of broader support, economies of scale, and a better funding base.

- As extensions to Z39.50 are developed to support image databases, NAL should participate in these standards development activities.

- NAL should carefully monitor other developments in standards related to imaging, including the work on X Windows. In some cases NAL may wish actually to participate or to have formal input into the development of these standards.

- NAL should monitor developments in distributed database and data-distribution protocols and consider participating in these developments as appropriate.

6.3 Projects and Prototypes

I believe that two projects should be undertaken as soon as feasible. Definition of further efforts will need to await the resolution of some of the broader policy issues discussed in section 6.1.

6.3.1 Internet Connection

The current NATDP system installed at NAL runs TCP/IP and can be connected to the internet. I recommend that NAL pursue such connection as a high priority by installing a link to some other nearby organization that is already part of the internet. Among the possible sites are NASA Greenbelt and the University of Maryland.

Hardware for such a connection should cost less that $20,000. There will be some ongoing charges for the telecommunications link; these will depend on bandwidth and on the location of the institution to which NAL is linking.

Once such a connection is established, NAL can conduct experiments with a number of other institutions that are already connected to the internet, including several participants in the Text

Digitizing Project. Suggested experiments include remote access to NAL databases, use of electronic mail for document requests and feedback on the published databases being tested in participant libraries, and network image transmission. Such experiments can be undertaken easily at little or no additional cost.

Termination of an internet connection at NAL also raises a variety of questions about local area networking at NAL's facility in Beltsville. As local area networking technology is deployed at the library, TCP/IP capability and the ability to support a gateway to the internet must be included in the LAN system requirements. Demand for access to resources on the internet may also accelerate demand for LAN deployment throughout the library.

6.3.2 Development of LAN Server Versions of CD-ROM Databases

NAL should work with one or more participant sites and one or more vendors of retrieval software to develop retrieval software that can be used to connect a workstation to a local area network at a participant site and that offers TELNET, Z39.50, and/or X Windows access to the database. One possible approach would be to implement different interfaces with different retrieval software packages. Since these capabilities are not generally available in retrieval software today, this project would essentially be a joint research and development project with one or more software vendors.

The costs of such an undertaking would depend to a great extent on the arrangements that could be made with the participating software vendors. Since the approach here represents a significant advance in the state of the art for CD-ROM retrieval software, it seems at least possible that the participating software vendors would underwrite a large portion of the costs. It may also be possible to obtain grant funding from sources such as NSF for such a project, since it would serve as an important prototype for many future databases.

TCP/IP/TELNET and X windows based software should be available now, at least in prototype. Z39.50 support is more complex, and could be included as a follow-on. Some specific systems we should investigate include:

- KnowledgeSet. They have announced both TCP/IP and X windows support on UNIX platforms.

- Personal Library Software. PLS is working on a X windows implementation.

- NYSERNET (The New York State Regional NSF Network). NYSERNET is working on a Z39.50 implementation running on a UNIX workstation.

Several sites, including Iowa State, the University of Hawaii and UC Davis have expressed interest in serving as test environments. One approach would be to "host" a different server on each campus LAN; since all these sites are part of the internet, sites could experiment with servers hosted at other sites across the internet.

NAL may also want to work with Iowa State, which is experimenting with CD-ROM network *disk* (as opposed to database) servers, which are currently available from Meridian Data and Online (among others). TCP/IP support is not yet available for these products, but Meridian is working on it and may deliver it in 1989.

6.4 Standards and Compatibility Considerations

Compatibility problems are beginning to appear between the bit-mapped displays being used on the NAL workstations and various CD-ROM retrieval packages. Basically, the problem is that bit-mapped displays are not standardized, and retrieval software tends to be overly hardware dependent. To facilitate the introduction of future hardware and software products into the Text Digitizing Project, a cleaner interface between retrieval software and display hardware (and software drivers) is required. X Windows technology is one way that a more formal interface can be established.

I suggest that in procuring the next generation of display devices and retrieval software, NAL give serious consideration to requiring that any display device be supported by an X Windows server on the workstation, and that any retrieval package write to this display by operating as an X Windows client rather than doing so directly. In a stand-alone configuration, the X client could communicate with the X server though local interprocess communications rather than over a network.

Such an approach would not only accomplish the required separation between display hardware and retrieval software, it would also provide local network access to the database through X Windows in a very simple extension of the stand-alone system technology.

To preserve the investment in the existing bit-mapped display hardware base, NAL should explore the possibility of implementing an X server on the hardware currently being supplied by SAIC. These display workstations could then be used as part of LAN access experiment in conjunction with a database server that supported X. With the addition of a LAN interface card, the current ASCII workstations could be used for TELNET access to a text-based TCP/IP database server.

Appendix 7

Vendor Names and Addresses

Science Applications International Corporation
10260 Campus Point Drive
San Diego, California 92121
(619) 546-6000

Saztec International, Inc.
6700 Corporate Drive, Suite 100
Kansas City, Missouri 64120
(818) 483-6900

Textware Corporation
P.O. Box 3267
Park City, Utah 84060
(801) 645-9600

Knowledge Access International Inc.
2685 Marine Way, Suite 1305
Mountain View, California 94043
(415) 969-0606

Personal Library Software, Inc.
2400 Research Blvd., Suite 350
Rockville, Maryland 20850
(301) 990-1155

Appendix 8

Summary of Phase III

Years 1 and 2

North Carolina State University

The Libraries

Box 7111
Raleigh
North Carolina 27695-7111
(919) 737-2843

TELEFAX:
Director's Office (919) 737-3628
Interlibrary Center (919) 737-7854
Photocopy Services (919) 737-7098

The North Carolina State University Libraries and the National Agricultural Library Joint Project on Transmission of Digitized Images: Improving Access to Agricultural Information

The National Agricultural Library (NAL), the North Carolina State University (NCSU) Libraries, and the NCSU Computing Center are collaborators in an ongoing research and demonstration project to discover and explore the issues involved in an NSFnet/Internet-based document delivery system for library materials. The project uses scanned images of documents to generate highly detailed, machine-readable text, and transmits those materials through the data transmission capabilities of the NSFnet/Internet to computers located in libraries, in research areas, on scholars' desks, at agricultural research stations and extension offices, etc. associated with a representative subset of land grant universities. An initial pilot project has established a data communications link between the NAL and the NCSU Libraries through SURAnet, the Southeastern Universities Research Association Network, one of the component networks of the interconnected group of computer networks known as the Internet. The pilot study has successfully transmitted digitized document pages between computers at NAL and computers at NCSU via the Internet. Images of documents can be delivered directly to a researcher's computer, placed on diskette, or printed. Printed copies of transmitted materials confirm the marked superiority of this delivery process over telefacsimile for reproducing graphical and photographic information important to scientific publications, such as line graphs and mathematical formulae. The pilot study also has demonstrated that by adhering to standards applied to data format and data communications, such as the TCP/IP protocol suite used by computers in the Internet, and by utilizing off-the-shelf equipment, it is possible to integrate dissimilar computing environments in libraries and research areas.

The NCSU Libraries has designed an expanded, full-scale implementation test of the pilot project system to be conducted during fiscal years 1990 through 1992 . The expanded system will place off-the-shelf, graphics-capable, networked desktop computers, scanners, and laser printers at up to eight land-grant university libraries and will link those computers to the national Internet. Participating libraries will use the equipment to digitize, transmit, and receive documents, including all text and illustrations, requested by agricultural researchers. The documents will be transmitted through the Internet to recipient libraries. Transmission of images and electronic mail exchange will be through the standard TELNET/FTP capabilities of the TCP/IP protocol suite.

The project management comprises a team of six persons from NCSU Libraries and the NCSU Academic Computing, including the Director of Libraries, the Libraries' Assistant Directors for Library Systems and Public Services, the Associate Provost for Academic Computing, the Director of the Computing Center, and a project manager. The project team, working closely with the NAL and its National Agricultural Text Digitizing Project Advisory Panel, will employ an established nomination and review mechanism to solicit applications and select up to eight participants for the expanded project. Three participants have already been designated: the NCSU Libraries, the NAL, and one 1890 land-grant institution representing the historically black colleges and universities, North Carolina

Agricultural and Technical University at Greensboro, N. C. In addition, the project may include international participation; a number of international agencies, as well as one country which has a cooperative agricultural program with NCSU, have expressed interest in participating in the expanded project.

The issue of copyright plays a central and crucial role in developing network-based document delivery systems. The research and demonstration project will address copyright in three ways. First, the project will transmit and receive only materials which fall either under the fair-use exemption of the copyright law or which are not copyrighted. Second, the project will conduct extensive research into the copyright issue specifically as it pertains to the transmission of digitized documents. The study will call upon local and national experts in copyright and explore related copyright areas. The results of the research will be published, and project staff plan to develop an ARL (Association of Research Libraries) SPEC kit that will summarize the current state of affairs regarding the copyright issue as it pertains to the transmission of digitized documents and provide descriptive materials on policies, procedures, and management techniques. Third, the NAL will continue its ongoing negotiations with academic publishers to provide copyright clearance for distribution of copyrighted research literature. These negotiations have been successful in at least one instance and are expected to yield further successes.

The agricultural industry in the U.S. is massive in expenditures and scope, important to every state in the Union and accounting for 14.9% of the gross national product, a total of $727 billion in 1989 alone. This enormous economic contribution is supported by an extensive research and development structure. The U.S. government alone spent over $2.1 billion for agricultural research services in 1989, and private sector expenditures on research and development mounted into the hundreds of millions of dollars. This research is widely distributed among federal and private laboratories, agricultural research stations located throughout the U.S, and colleges and land-grant universities. The size and distributed nature of the agricultural research effort underscores the intense demand to improve document delivery. The requirement for prompt delivery of library materials to researchers, combined with the evolving sophistication of researchers in the application of computers to research problems and the increasing demands to integrate library support into scientists' research functions, provides strong incentive to develop network-based systems for delivery of documents that overcome the pronounced drawbacks of existing technical and procedural mechanisms, such as facsimile and postal delivery of photocopies.

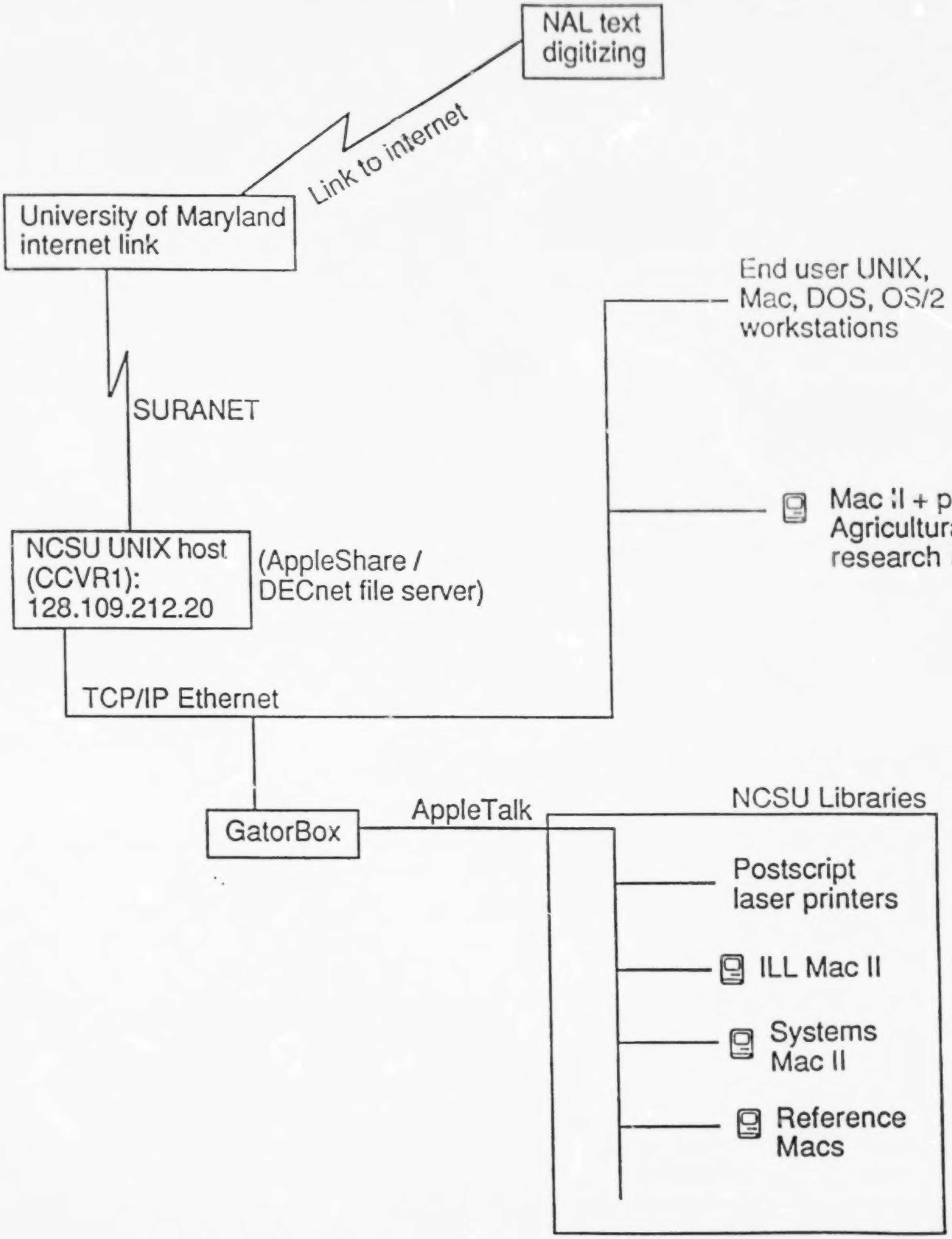
The proposed project is a large-scale demonstration system that will utilize high-speed networks for routine delivery of a large class of materials. Its size and scope will lay the groundwork for the development of a full production system by the mid-1990's that will use national Internet facilities for document delivery among the entire land-grant university community in the U.S., as well as to other Internet sites and to the federal and international agricultural research community. In this respect, an important objective of the expanded demonstration system is to strengthen the network infrastructure employed by the agricultural science research community and encourage its application to information dissemination and delivery, in anticipation of prominent role network-based systems are likely to play in this decade. The project will provide state-of-the-art computer and network technologies that will help develop the expertise of librarians and information workers in using and managing network technologies and network-based services.

For additional information, please contact either:

Susan K. Nutter
Director of Libraries, NCSU
SKNDHHL@NCSUVM.BITNET
SKNDHHL@NCSUVM.NCSU.EDU

John E. Ulmschneider
Assistant Director for Library Systems, NCSU
JEUDHHL@NCSUVM.BITNET
JEUDHHL@NCSUVM.NCSU.EDU

Digitized Text Transfer
Logical Network Schematic


NAL text digitizing
Link to internet
University of Maryland internet link
SURANET
NCSU UNIX host (CCVR1): 128.109.212.20
(AppleShare / DECnet file server)
End user UNIX, Mac, DOS, OS/2 workstations
Mac II + printer: Agricultural research area
TCP/IP Ethernet
GatorBox
AppleTalk
NCSU Libraries
Postscript laser printers
ILL Mac II
Systems Mac II
Reference Macs

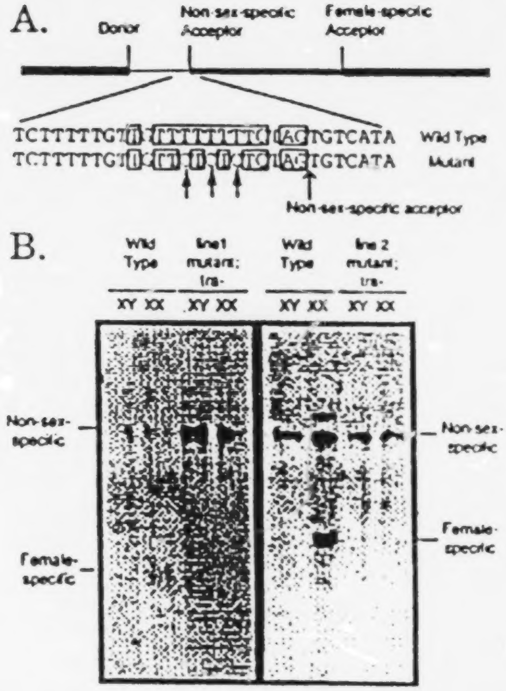


Figure 5. RNAase Mapping of the Altered Non-Sex-Specific Splice Site Mutant

(A) Schematic diagram of the alternatively spliced region of *tra*, with the sequences of the wild-type and a mutant non-sex-specific splice site. Boxes around the sequences indicate regions of sequence identity between *tra* and *Sxl* as shown in Figure 1. The mutations (marked with arrows) disrupt the central portion of the region of sequence identity but do not change the size of the polypyrimidine region and retain all elements of a consensus 3' splice site.

(B) RNAs from wild-type or *tra*⁻ flies containing the mutant 3' splice site were used for RNAase protection analysis. The probe used is homologous to the mutant RNAs but, because of G-U base pairing, will pair with unspliced RNA from both wild-type and mutant genes. Splicing to the non-sex-specific splice site leads to protected fragments of the same size for mutant and wild-type genes, as indicated. Splicing to the female site also leads to protected fragments of the same size for mutant and wild-type genes. The band in the wild-type female lanes that is ~75 nucleotides larger than the non-sex-specific RNA is highly female biased in its appearance and is of a size consistent with protection by unspliced RNA. RNAs from two independent lines of mutant flies are shown. Among the G418^R flies, *tra*⁻ flies were identified on the basis of mutant eye color and bristle morphology resulting from markers closely linked to *tra*. XX and XY flies were distinguished by the presence of a dominant Y chromosome marker, *Bar*.

It should be noted that results such as these would also be predicted by the production of a non-sex-specific 3' splice site which was more efficient such that it could not be repressed or out-competed even by an activated female 3' splice site. We take our reproducible result that the level of constitutive use of the female splice site is actually increased in multiple lines carrying this mutant gene to be evidence against this hypothesis. The small degree of apparent loss of efficiency of the non-sex-specific splice site makes the experimental results clearer (i.e., lack of female-specific activation is easier to see), but makes it harder to definitively rule out the possibility that the mutant 3' splice site is in some sense "better."

Wild-Type Females Accumulate Unspliced RNA

An examination of Figures 4B and 5B shows the presence in the wild-type female lanes of an additional band of RNAase-resistant material at a position 70–80 nucleotides larger than the band resulting from protection by RNA spliced in the non-sex-specific manner. This is most clearly seen in the right-hand gels of Figures 4B and 5B, but it is also present, though partially obscured by background bands, in the left-hand gels of these figures. The size of this RNA is exactly the size expected of an RNA protected from digestion by unspliced *tra* RNA. From Figure 5B and a number of other gels it is clear that this RNA is highly female biased and represents a much larger fraction of the *tra* protection in RNA derived from females than in RNA derived from males. Similar results have been obtained by S1 nuclease mapping (Boggs et al., 1987; R. T. Boggs and M. M., unpublished data). Although the blockage model does not require the female-specific accumulation of unspliced *tra* RNAs, this result is fully consistent with a female-specific factor binding to the non-sex-specific 3' splice site (thereby lowering its usage) without directly increasing the efficiency of the female-specific 3' splice site. Female-specific accumulation of unspliced RNA could also be explained by versions of activation models in which either the process of deciding between two splice sites or the commitment to using the female splice site slows the overall rate of splicing and thereby increases the steady-state level of unspliced RNA.

As can be seen in Figure 5B, the base substitution mutant does not accumulate unspliced RNA in a sex-specific manner. Thus this mutant does not respond to *Sxl* in the production of the female RNA and does not respond to *Sxl* in the accumulation of unspliced precursor, two aspects of the female-specific regulation of *tra*.

Female-Specific Accumulation of Unspliced RNA Occurs in the Absence of the Female 3' Splice Site

To distinguish between alternative possibilities for the accumulation of unspliced RNA, we have created flies that carry a mutant *tra* gene (ΔM4) deleted for the female splice site. This deletion removes 18 bases 5' of the splice site and 6 bases 3' of the splice site and replaces them with a 10 base EcoRI linker. The blockage model hypothesizes that failure to splice is a result of an interaction between the non-sex-specific splice site and a female-specific factor without regard for other potential splice sites and predicts that this mutant should show a female-specific buildup of unspliced RNA. The activation model posits that the unspliced RNA found in wild-type females results from competition between the non-sex-specific 3' splice site and an activated female splice site, or that the unspliced RNA results from a slow step in the initial utilization of the female splice site. If it is the utilization of an activated female splice site or the competition between the two splice sites that results in unspliced RNA, then deletion of the female site should result in a failure to accumulate unspliced RNA in a female-specific manner.

BEST COPY AVAILABLE

North Carolina State University

The Libraries

Box 7111
Raleigh
North Carolina 27695-7111
(919) 515-2843

TELEFAX:
Director's Office (919) 515-3628
Interlibrary Center (919) 515-7854
Photocopy Services (919) 515-7098

NCSU Digitized Document Delivery Project

Electronic Document Delivery Service
A Pilot Service

The Natural Resources Library (NRL) has been selected to participate in the first phase of the Libraries' pilot network service entitled, ***Electronic Document Delivery Service*** (EDDS). The EDD Service will allow library researchers in the College of Forest Resources and the Department of Marine, Earth, and Atmospheric Sciences to place and receive their interlibrary loan journal article requests electronically via the campus telecommunications network. The pilot service is scheduled to begin December 9, 1991 and run through July 31, 1992. (See Electronic Document Delivery Service: A Pilot Service Logical Schematic.)

The EDD Service is part of a larger research initiative conducted by The Libraries entitled, ***NCSU Digitized Document Transmission Project (DDTP)***. The Libraries in cooperation with the NCSU Computing Center, the National Agricultural Library (NAL) and twelve land-grant libraries seek to demonstrate the technical feasibility of delivering library materials directly to the researcher in a timely, value-added, machine-readable form. The current project builds upon the findings of an earlier demonstration study on the transmission of digitized images conducted from April 1989 through September 1990 by the Libraries, the NCSU Computing Center and the NAL.

Using off-the-shelf, graphics-capable, networked hardware platforms and commercially available software, the fourteen participating libraries digitize documents using commercially available scanners and capture the image in Tagged Image File Format (TIFF), a non-proprietary machine-independent data format. The transmission of images is made possible through the standard TELNET/FTP capabilities of the TCP/IP protocol suite. Scanned image files are compressed for transmission using low-cost compression programs and can be decompressed by researchers using readily available shareware software. The project has already established that print copies of digitized documents rival high quality photocopies; the machine-readable images can be readily imported into text or graphics packages.

The EDD Service for the direct delivery of digitized journal articles consists of two components: electronic mail messaging and direct delivery/pickup of the digitized documents. The messaging system allows researchers to place their journal article requests via electronic-mail to project staff in the Interlibrary Center (ILC). Researchers may either enter the citation in free-form or use an electronic request template provided by the Service, available on the campus fileserver. The direct delivery component allows the researcher to retrieve the digitized document(s) from the central campus fileserver via the campus telecommunications network.

From the researcher's perspective, the EDD Service automates the entire process of requesting and receiving materials retrieved from other library's collections.

- The patron downloads a bibliographic citation from a library catalog accessible via the Internet or from an electronic database, then
- sends the request via electronic-mail to the ILL.
- ILL verifies and transmits the request via the OCLC ILL subsystem to a participating NCSU DDTP site.
- The lending library scans and transmits the digitized document to the designated campus fileserver at the borrowing library.
- An electronic mail notice is automatically issued by the fileserver to the patron informing him/her that the digitized journal article is available and may be retrieved via the FTP capabilities of the campus network.

The service is not limited solely to the delivery of journal articles; any type of library information that can be captured in digital form or that already exists in digital form can be delivered over the Internet and across campus networks to the researcher.

Presently, the pilot EDD Service is available only to researchers in the College of Forest Resources and the Department of Marine, Earth, and Atmospheric Sciences. Researchers interested in using the service will first need to register at the NRL . Upon registering, researchers will receive an information packet describing how to use the service. The packet contains shareware software for decompressing the files. The researcher is responsible for registering shareware software. The project staff has developed a custom application that will allow for the automatic printing of the images should a researcher not have access a graphic program such as Pagemaker, Canvas, SuperPaint or PC Paintbrush to view the image. The service is designed to support both the Macintosh and DOS environments. Patrons without the required equipment and/or connectivity will have the option of requesting and receiving digitized journal articles via the NRL staff from the project workstation located in the NRL.

Project staff will draw on the collections of the fourteen participating land-grant libraries to fill the requests for documents not available from the Libraries' collections. Requests that cannot be filled through the EDD Service will be processed immediately via regular ILL channels (if desired).

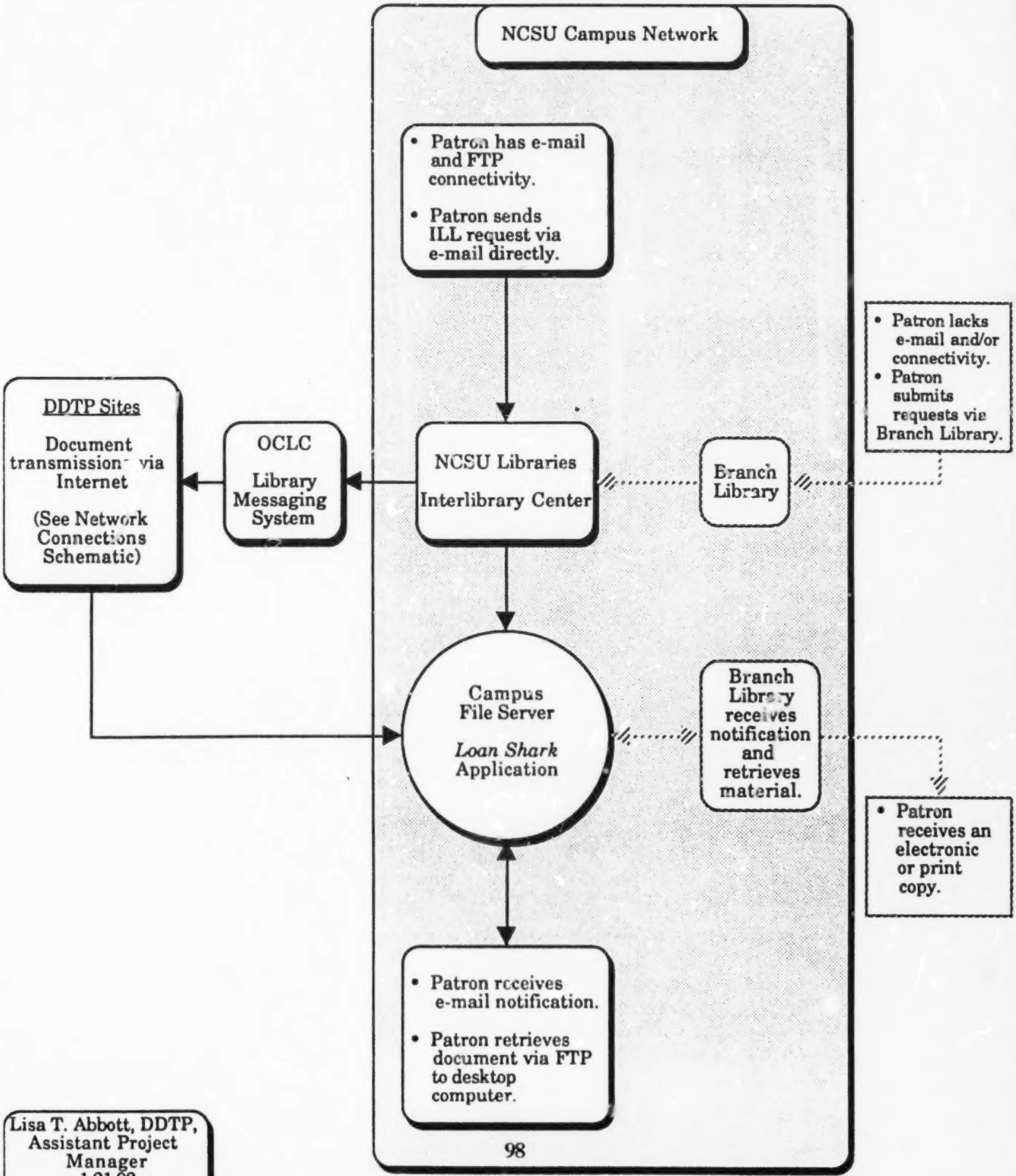
The twelve land-grant institutions participating with NCSU and NAL in the project are: Clemson University, University of Delaware, Iowa State University, University of Maryland at College Park, Michigan State University, University of Minnesota, North Carolina Agricultural and Technical State University, The Ohio State University, Pennsylvania State University, Utah State University, Virginia Polytechnic Institute & State University, and Washington State University. (See attached Network Connections Schematic.)

Lisa Abbott, DDTP Assistant Manager, is overseeing the EDD Service working closely with the staff in ILL; Carolyn Argentati, Head, Natural Resources Library; and Sam Moore of the Computing Center to implement and refine the EDDS model. This component of the DDTP has been made possible through the Apple Library of Tomorrow (ALOT) equipment grant. The NCSU DDTP is funded in part by a U.S. Department of Education Title II-D Research and Demonstration grant and participating institutions.

For more information contact:
Lisa T. Abbott, DDTP Assistant Manager
(919) 515-3708
Lisa_Abbott@NCSU.EDU

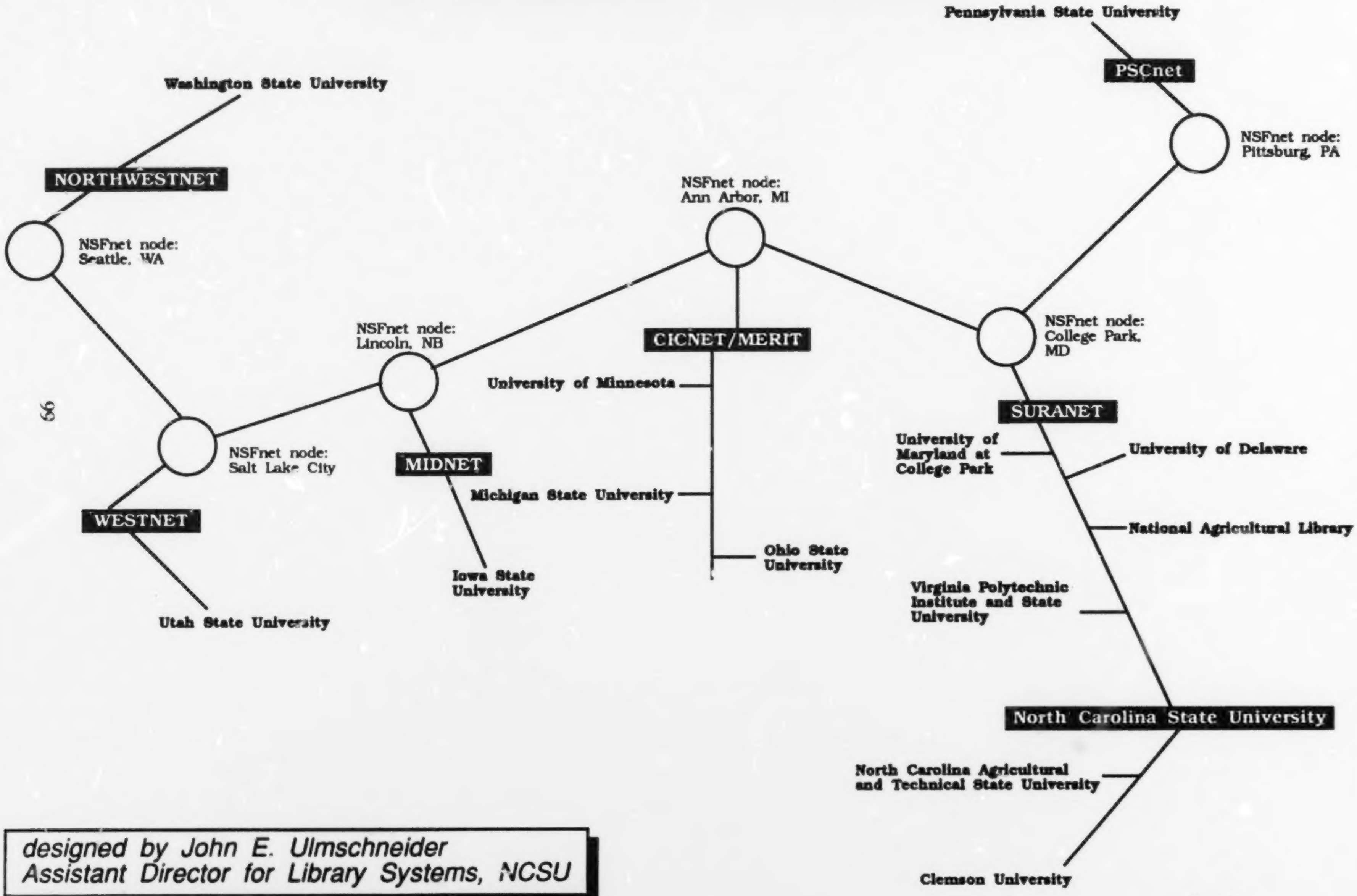
1-8-92

NCSU Digitized Document Transmission Project
Electronic Document Delivery Service: A Pilot Service
Logical Schematic


NCSU Campus Network
• Patron has e-mail and FTP connectivity.
• Patron sends ILL request via e-mail directly.
DDTP Sites
Document transmission via Internet
(See Network Connections Schematic)
OCLC
Library Messaging System
NCSU Libraries
Interlibrary Center
Branch Library
• Patron lacks e-mail and/or connectivity.
• Patron submits requests via Branch Library.
Campus File Server
Loan Shark Application
Branch Library receives notification and retrieves material.
• Patron receives an electronic or print copy.
• Patron receives e-mail notification.
• Patron retrieves document via FTP to desktop computer.

Lisa T. Abbott, DDTP,
Assistant Project
Manager
1-21-92

NCSU Digitized Document Transmission Project
Network Connections: Logical Schematic
June 1991


Pennsylvania State University
PSCnet
NSFnet node: Pittsburg, PA
Washington State University
NORTHWESTNET
NSFnet node: Seattle, WA
NSFnet node: Ann Arbor, MI
NSFnet node: Lincoln, NB
CICNET/MERIT
NSFnet node: College Park, MD
University of Minnesota
SURANET
NSFnet node: Salt Lake City
MIDNET
University of Maryland at College Park
University of Delaware
Michigan State University
WESTNET
National Agricultural Library
Ohio State University
Iowa State University
Virginia Polytechnic Institute and State University
Utah State University
North Carolina State University
North Carolina Agricultural and Technical State University
Clemson University

designed by John E. Ulmschneider
Assistant Director for Library Systems, NCSU

Appendix 9

Bibliography of Articles about the NATDP

André, Pamela Q.J. "The Library as an Information Producer in a Networked Environment: The National Agricultural Library Experience." In *Proceedings of the ACRL New England Chapter's Spring Conference, March 1991*. Westport, Connecticutt: Meckler. In press.

André, Pamela Q.J. "Optical Technology: New Ways of Providing Information to the End User." *Quarterly Bulletin of the International Association of Agricultural Information Specialists* 36, nos. 1-2 (1991): 136-38.

André, Pamela Q.J. "Towards the Electronic Library: The National Agricultural Library Experience with CD-ROM Technology." In *CD-ROM for International Development, Proceedings of a Workshop, Geneva, 16-18 December 1991*, 1-8. (Netherlands): Technical Centre for Agricultural and Rural Cooperation, 1992.

André, Pamela Q.J. and Eaton, Nancy L. "National Agricultural Text Digitizing Project." *Library Hi Tech* 6, no. 3 (Consecutive Issue 23, 1988): 61-66.

André, Pamela Q.J., Eaton, Nancy, and Zidar, Judith. "Scanning and Digitizing Technology Employed in the National Agricultural Text Digitizing Project." In *Proceedings of the Conference on Application of Scanning Methodologies in Libraries*, ed. D.L. Blamberg, C.L. Dowling, C.V. Weston, 61-73. Beltsville, Maryland: National Agricultural Library, 1989.

Eaton, Nancy L. and André, Pamela Q.J. "National Agricultural Text Digitizing Project." In *The Electronic Library: Linking People, Information, and Technology*, 19-21. RASD Occasional Papers no. 14. Chicago: American Library Association, Reference and Adult Services Division, 1992.

Joy, Albert H., Eaton, Nancy L., and Goins, Rodney K. "Access to Canadian Government Publications on Acid Rain: The University of Vermont's HEA Title II-C Grant." *Government Publications Review* 16 (January/February 1989): 31-39.

Zidar, Judith. "Optical Scanning and Test Recognition: Operating on In-House System." *Quarterly Bulletin of the International Association of Agricultural Information Specialists* 37, nos. 1-2 (1992): 65-69.